Title
Machine learning predicts new anti-CRISPR proteins.

Journal
Nucleic acids research, 48(9)

ISSN
0305-1048

Authors
Eitzinger, Simon
Asif, Amina
Watters, Kyle E
et al.

Publication Date
2020-05-01

DOI
10.1093/nar/gkaa219

Peer reviewed

Machine Learning Predicts New Anti-CRISPR Proteins

Simon Eitzinger[1†], Amina Asif[2†], Kyle E. Watters[1†], Anthony T. Iavarone[3], Gavin J. Knott[1], Jennifer A. Doudna[1,4-8*], and Fayyaz ul Amir Afsar Minhas[2,9*]

[1]Department of Molecular and Cell Biology, University of California Berkeley, Berkeley, CA 94720, USA

[2]Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences (PIEAS), PO Nilore, Islamabad, Pakistan.

[3]QB3/Chemistry Mass Spectrometry Facility, University of California, Berkeley, Berkeley, CA, USA

[4]Department of Chemistry, University of California Berkeley, Berkeley, CA, 94720, USA

[5]Innovative Genomics Initiative, University of California Berkeley, Berkeley, CA, 94720, USA

[6]Center for RNA Systems Biology, University of California Berkeley, Berkeley, CA, 94720, USA

[7]Howard Hughes Medical Institute, University of California Berkeley, Berkeley, CA, 94720, USA

[8]Molecular Biophysics & Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA

[9]Department of Computer Science, University of Warwick, Coventry, CV4 7AL, UK

† These authors contributed equally to this work.

*To whom correspondence should be addressed: doudna@berkeley.edu, fayyaz.minhas14@alumni.colostate.edu

ABSTRACT

The increasing use of CRISPR-Cas9 in medicine, agriculture, and synthetic biology has accelerated the drive to discover new CRISPR-Cas inhibitors as potential mechanisms of control for gene editing applications. Many anti-CRISPRs have been found that inhibit the CRISPR-Cas adaptive immune system. However, comparing all currently known anti-CRISPRs does not reveal a shared set of properties for facile bioinformatic identification of new anti-CRISPR families. Here, we describe AcRanker, a machine learning based method to aid direct identification of new potential anti-CRISPRs using only protein sequence information. Using a training set of known anti-CRISPRs, we built a model based on XGBoost ranking. We then applied AcRanker to predict candidate anti-CRISPRs from predicted prophage regions within self-targeting bacterial genomes and discovered two previously unknown anti-CRISPRs: AcrllA20 (ML1) and AcrIIA21 (ML8). We show that AcrIIA20 strongly inhibits *Streptococcus iniae* Cas9 (SinCas9) and weakly inhibits *Streptococcus pyogenes* Cas9 (SpyCas9). We also show that AcrIIA21 inhibits SpyCas9, *Streptococcus aureus* Cas9 (SauCas9) and SinCas9 with low potency. The addition of AcRanker to the anti-CRISPR discovery toolkit allows researchers to directly rank potential anti-CRISPR candidate genes for increased speed in testing and validation of new anti-CRISPRs. A web server implementation for AcRanker is available online at http://acranker.pythonanywhere.com/.

INTRODUCTION

CRISPR-Cas systems use a combination of genetic memory and highly specific nucleases to form a powerful adaptive defense mechanism in bacteria and archaea (1–4). Due to their high degree of sequence specificity, CRISPR-Cas systems have been adapted for use as programmable DNA or RNA editing tools with novel applications in biotechnology, diagnostics, medicine, agriculture, and more (5–9). In 2013, the first anti-CRISPR proteins (Acrs) were discovered in *Pseudomonas aeruginosa* phages able to inhibit the CRISPR-Cas system (10). Since then, Acrs able to inhibit a wide variety of different CRISPR subtypes have been found (10-28).

Multiple methods for identifying Acrs include screening for phages that escape CRISPR targeting (10, 19–23), guilt-by-association studies (12, 17, 24, 25, 28), identification and screening of genomes containing self-targeting CRISPR arrays (11–13, 24), and metagenome DNA screening for inhibition activity (26, 27). Of these approaches, the 'guilt-by-association' search strategy is one of the most effective and direct, but it requires a known Acr to serve as a seed for the search. Thus, the discovery of one new validated Acr can lead to bioinformatic identification of others, as many Acrs have been discovered to be encoded in close physical proximity to each other, typically co-occurring in the same transcript with other Acrs or anti-CRISPR associated (*aca*) genes (12, 17, 28). Screening approaches are particularly useful in this regard, as they can potentially identify new Acr families.

Identification of self-targeting CRISPR arrays can also help in predicting new Acr families. Typically, a CRISPR array with a spacer targeting the host genome (self-targeting) is lethal to the cell (29). However, if a mobile genetic element (MGE) present in the cell carries *acr* genes, the CRISPR-Cas system could be inhibited, and this may allow a cell with a self-targeting array to survive. To find new Acrs, genomes containing self-targeting arrays are identified through bioinformatic methods, and the MGEs within are screened for anti-CRISPR activity, eventually narrowing down to individual proteins (11–13, 24). Screens based on self-targeting also benefit from the knowledge of the exact CRISPR system that an inhibitor potentially exists for, as opposed to broad (meta-)genomic screens where a specific Cas protein has to be selected to screen against. Both types of screening additionally benefit from not requiring the prediction of a transcriptome or proteome that bioinformatic methods depend on, where incorrect annotations could lead to missed *acr* genes (24).

However, a weakness of all of these methods is that they are unable to predict *a priori* whether a gene may be an Acr, largely because Acr proteins do not share high sequence similarity or mechanisms of action (14, 16, 30–36). One theory to explain the high diversity of Acrs is the rapid mutation rate of the mobile genetic elements they are found in and the need to evolve with the co-evolving CRISPR-Cas systems trying to evade anti-CRISPR activity. Due to the relatively small size of most Acrs and their broad sequence diversity, simple sequence comparison methods for searching anti-CRISPR proteins are not expected to be effective. In this work, we report the development of AcRanker, a machine learning based method for direct identification of anti-CRISPR proteins. Using only amino acid composition features, AcRanker ranks a set of candidate proteins on their likelihood of being an anti-CRISPR protein. A rigorous cross-validation of the proposed scheme shows known Acrs are highly ranked out of proteomes. We then use AcRanker to predict 10 new candidate Acrs from proteomes of bacteria with self-targeting CRISPR arrays and biochemically validate three of them. Our machine learning approach presents a new tool to directly identify potential Acrs for biochemical validation using protein sequence alone.

MATERIALS AND METHODS

Data collection and preprocessing

To model the task of anti-CRISPR protein identification as a machine learning problem, a dataset consisting of examples from both positive (anti-CRISPR) and negative (non-anti-CRISPR) classes was needed. We collected anti-CRISPR information for proteins from the Anti-CRISPRdb (37). At the time the work was initiated, the database contained information for 432 anti-CRISPR proteins. In order to ensure that the machine learning model generalizes well to protein sequences that do not share high sequence similarity to known anti-CRISPR proteins, a 40% sequence identity threshold is used (38). The use of a 40% identity threshold represents a boundary where proteins above this threshold are likely to share the same structure and possibly function (39), thus providing a comprise between ensuring non-redundancy of the train and test datasets while retaining enough training examples for cross-validation. We used CD-HIT (40) to identify a non-redundant set (at the 40% sequence similarity threshold) of 20 experimentally verified Acrs (Table S1). These proteins belong to different Acr classes: 12 of the proteins are active against subtype I-F

CRISPR Cas systems, four against I-E, and four against II-A (10, 13, 17, 20, 22). This set constitutes the positive class of our dataset. We downloaded the complete proteomes of source species to which each of these proteins belong. Within these proteomes, any protein with 40% or higher sequence similarity with any protein in the set of known anti-CRISPR proteins was removed, and the remaining proteins were used to construct the negative dataset. For independent testing of the method, a dataset comprising 20 known Acrs separate from the training set (11–13, 21, 24, 26, 28, 41) was used (Table S2). The Acrs belonging to the test set were chosen to cover the wide variety of known Acr mechanisms and sequences (42), while mainly consisting of the three subtypes the model was trained on. Source proteomes for all these proteins were downloaded, based on open reading frame predictions on the NCBI database.

Feature Extraction

In line with existing machine learning based protein function prediction techniques, we used sequence features (43) based on amino acid composition and grouped dimer and trimer frequency counts (44). For this purpose, amino acids are first grouped into seven classes based on their physicochemical properties (44) (Table S3) and the frequency counts of all possible groups labeled as dimers and trimers in a given protein sequence are used in conjunction with amino acid composition. All three types of features (amino acid composition, di- and tri- meric frequency counts) are normalized to unit norm resulting in a $20 + 7^2 + 7^3 = 412$-dimensional feature vector representation for a given protein sequence (45, 46).

Machine learning model

The underlying machine learning model for AcRanker has been built using EXtreme Gradient Boosting (XGBoost) (47). In machine learning, boosting is a technique in which multiple weak classifiers are combined to produce a strong classifier (47). XGBoost is a tree-based method (47) that uses boosting in an end-to-end fashion, i.e., every next tree tries to minimize the error produced by its predecessor. XGBoost has been shown to be a fast and scalable learning algorithm and has been widely used in many machine learning applications (47).

In this work, we have used XGBoost as a pairwise ranking model to rank constituent proteins in a given proteome in descending order of their expected Acr behavior. The XGBoost model is trained in a proteome-

specific manner to produce higher scores for known anti-CRISPR proteins as compared to non-anti-CRISPR proteins in a given proteome. In comparison to conventional XGBoost classification, the pairwise ranking model performed better in terms of correctly identifying known anti-CRISPR proteins in test proteomes in cross-validation (Table S4). Specifically, given a set of training proteomes S each with one or more known anti-CRISPR proteins, our objective is to obtain an XGBoost predictor $f(\boldsymbol{x};\theta)$ with learnable parameters θ that generates a prediction score for a given protein sequence represented in terms of its feature vector $\boldsymbol{x}$. In proteome-specific training, we require the model to learn optimal parameters θ^* such that the score $f(\boldsymbol{p};\theta^*)$ for a positive example $\boldsymbol{p}$ (known Anti-CRISPR protein) should be higher than $f(\boldsymbol{n};\theta^*)$ for all negative examples $\boldsymbol{n}$ (non-Anti-CRISPR proteins) within the same proteome. The hyperparameters of the learning model are selected through cross validation and optimal results are obtained with: number of estimators set at 120, learning rate of 0.1, subsampling of 0.6, and maximum tree depth of 3.

Performance Evaluation

To evaluate the performance of the machine learning model, we have performed leave-one-proteome-out cross-validation as well as validation over an independent test set. In a single fold of leave-one-proteome-out cross-validation, we set aside the source proteome of a given anti-CRISPR protein for testing and train on all other proteomes. To ensure an unbiased evaluation, all sequences in the training set with a sequence identity of 40% or higher with any test protein or among themselves are removed from the training set. Furthermore, all proteins in the test set with more than 40% sequence identity with known anti-CRISPR proteins in the training set are also removed. This ensures that there is only one known anti-CRISPR protein in the test set in a single fold. The XGBoost ranking model is then trained and the prediction scores for all proteins in the test set are computed. Ideally, the known anti-CRISPR protein in the proteome should score the highest across all proteins in the given test proteome. This process is then repeated for all proteomes in our dataset. The rank of the known anti-CRISPR protein in its source proteome is used as a performance metric.

In bacteria, Acrs are usually located within prophage regions (13, 48). Based on this premise, in another experiment for model evaluation, we passed only the proteins found within prophage regions to the model.

To identify the prophage regions for a given bacterial proteome we used PHASTER (PHAge Search Tool Enhanced Release) web server (49) which accepts a bacterial genome and annotates prophage regions in it. The decision scores are computed for all phage proteins identified by PHASTER in the test proteome.

To help assess AcRanker's performance during leave-one-out cross-validation, BLAST (Basic Local Alignment Search Tool) (50) similarity was used to set a minimum performance expectation. For each protein in a given test proteome, we compute BLASTp scores (with default parameters) with the set of known Acrs (excluding the tested protein) and rank proteins in the increasing order of the respective e-values.

For independent validation, the ranking based XGBoost model trained over sequence features for all 20 source proteomes (Table S1) has been tested for recently discovered Acrs (Table S2) that are not part of our training set. The rank of known Acr in its corresponding proteome was computed. Here again, we evaluated the model for both the complete proteome of the organism and the respective MGE subset identified by PHASTER.

AcRanker Webserver

A webserver implementation of AcRanker is publicly available at http://acranker.pythonanywhere.com/. The webserver accepts a proteome file in FASTA format and returns a ranked list of proteins. The Python code for the webserver implementation is available at the URL: https://github.com/amina01/AcRanker.

Acr candidate selection

Self-Targeting Spacer Searcher (STSS; https://github.com/kew222/Self-Targeting-Spacer-Searcher) (11) was run with default parameters using 'Streptococcus' as a search term for the NCBI genomes database, which returned a list of all self-targets found in those genomes. Whether known *acr* genes were present in each of the self-targeting genomes was checked using a simple blastp search using default parameters with the Acr proteins stored within STSS. Twenty self-targeting genomes that contained at least one self-target with a 3′-NRG PAM were chosen for further analysis with AcRanker. Prophage regions with each genome were predicted using PHASTER (49). Then proteins found across all of the prophage regions predicted in a given genome were ranked with AcRanker.

To select individual gene candidates for synthesis and biochemical validation, the 10 highest ranked proteins from each genome were examined by visual inspection for a strong promoter, a strong ribosome binding site, and an intrinsic terminator. Promoters were searched for manually by looking for sequences closely matching the strong consensus promoter sequence TTGACA-17(+/-1)N-TATAAT upstream of the *acr* candidate gene, or any genes immediately preceding it. The presence of a strong ribosome binding site (resembling AGGAGG) near the start codon was similarly searched for and was required to be upstream of a gene candidate for selection. Last, given the nature of Acrs to be clustered together, genes neighboring the best candidates were also selected for further testing/validation and comprise part of the 10-member candidate test set.

Protein expression and purification

Each of the Acr candidates (Table S5) were cloned into a custom vector (pET-based expression vector) such that each protein was N-terminally tagged with a 10xHis sequence, superfolder GFP, and a tobacco etch virus (TEV) protease cleavage site, available on Addgene (#140995-141004). Each Cas effector (Table S6): *Acidaminococcus sp*. Cas12a (AsCas12a), *Streptococcus pyogenes* Cas9 (SpyCas9), *Staphylococcus aureus* Cas9 (SauCas9) and *Streptococcus iniae* Cas9 (SinCas9, Addgene #141076), were expressed as N-terminal MBP fusions. Proteins were produced and purified as previously described (33). Briefly, *E. coli* Rosetta2 (DE3) containing Acr or Cas9 expression plasmids were grown in Terrific Broth (100 µg/mL ampicillin) to an OD_{600} of 0.6-0.8, cooled on ice, induced with 0.5 mM isopropyl-b-D-thiogalactoside and incubated with shaking at 16°C for 16 h. Cells were harvested by centrifugation, resuspended in wash buffer (20 mM Tris-Cl (pH 7.5), 500 mM NaCl, 1 mM tris(2-carboxyethyl)phosphine (TCEP), 5% (v/v) glycerol) supplemented with 0.5 mM phenylmethanesulfonyl fluoride and cOmplete protease inhibitor (Roche), lysed by sonication, clarified by centrifugation and purified over Ni-NTA Superflow resin (Qiagen) in wash buffer supplemented with 10 mM (wash) or 300 mM imidazole (elution). Elution fractions were pooled and digested overnight with recombinantly expressed TEV protease while dialyzed against dialysis buffer (20 mM Tris-Cl (pH 7.5), 125 mM NaCl, 1 mM TCEP, 5% (v/v) glycerol) at 4°C. The cleaved proteins were loaded onto an MBP-Trap (GE Healthcare) upstream of a Heparin Hi-Trap (GE Healthcare) in the case of SpyCas9, SauCas9 and SinCas9. Depending on the pI, TEV digested Acrs

were loaded onto a Q (ML1, ML2, ML3, ML6, ML8, and ML10), heparin (ML4 and ML5), or SP (ML7 and ML9) Hi-Trap column. Proteins were eluted over a salt gradient (20 mM Tris-Cl (pH 7.5), 1 mM TCEP, 5% (v/v) glycerol, 125 mM – 1 M KCl). The eluted proteins were concentrated and loaded onto a Superdex S200 Increase 10/300 (GE Healthcare) for SpyCas9, SauCas9, SinCas9 or Superdex S75 Increase 10/300 (GE Healthcare) for all the Acr candidates and developed in gel filtration buffer (20 mM HEPES-K (pH 7.5), 200 mM KCl, 1 mM TCEP and 5% (v/v) glycerol). The absorbance at 280 nm was measured by Nanodrop and the concentration was determined using an extinction coefficient estimated based on the primary amino acid sequence of each protein. Purified proteins were concentrated to approximately 50 µM for Cas9 effectors and 100 µM for Acr candidates. Proteins were then snap-frozen in liquid nitrogen for storage at -80°C. Purity and integrity of proteins was assessed by 4-20% gradient SDS-PAGE (Coomassie blue staining, Figure S2A) and LC-MS (Figure S2B).

RNA preparation

All RNAs (Table S7) were transcribed *in vitro* using recombinant T7 RNA polymerase and purified by gel extraction as described previously (51). Briefly, 100 µg/mL T7 polymerase, 1 µg/mL pyrophosphatase (Roche), 800 units RNase inhibitor, 5 mM ATP, 5 mM CTP, 5 mM GTP, 5 mM UTP, 10 mM DTT, were incubated with DNA target in transcription buffer (30 mM Tris-Cl pH 8.1, 25 mM $MgCl_2$, 0.01% Triton X-100, 2 mM spermidine) and incubated overnight at 37**°**C. The reaction was quenched by adding 5 units RNase-free DNase (Promega). Transcription reactions were purified by 12.5% (v/v) urea-denaturing PAGE (0.5x Tris-borate-EDTA (TBE)) and ethanol precipitation.

***In vitro* cleavage assay**

In vitro cleavage assays were performed at 37°C in 1X cleavage buffer (20 mM Tris-HCl pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 1 mM DTT and 5% glycerol (v/v)) targeting a PCR amplified fragment of double-stranded DNA (Table S8). For all cleavage reactions, the sgRNA was first incubated at 95°C for 5 min and cooled down to room temperature. The Cas effectors (SpyCas9, SauCas9, AsCas12a at 100 nM and SinCas9 at 200 nM respectively) were incubated with each candidate Acr protein at 37°C for 10 min before the addition of sgRNA (SpyCas9, SauCas9, AsCas12a sgRNA at 160 nM and SinCas9 sgRNA at 320 nM respectively)

to form the RNP at 37°C for 10 min. The DNA cleavage reaction was then initiated with the addition of DNA target and reactions incubated for 30 min at 37°C before quenching in 1X quench buffer (5% glycerol, 0.2% SDS, 50 mM EDTA). Samples were then directly loaded to a 1% (w/v) agarose gel stained with SYBRGold (ThermoFisher) and imaged with a BioRad ChemiDoc.

Competition binding experiment

The reconstitution of the SinCas9-sgRNA-ML1 and SinCas9-sgRNA-AcrIIA2 complex was carried out as previously described (52). Briefly, purified SinCas9 and *in vitro* transcribed sgRNA were incubated in a 1:1.6 molar ratio at 37°C for 10 min to form the RNP. To form the inhibitor bound complexes, a 10-fold molar excess of AcrIIA20 (ML1) or AcrIIA2 were added and incubated with the RNP complex at 37°C for 10 min. For the competition binding experiment, a 10-fold molar excess of AcrIIA20 was first incubated with the RNP complex at 37°C before incubation with a 10-fold molar excess of AcrIIA2 at 37°C for 10 min. Each complex was then purified by analytical size-exclusion chromatography (Superdex S200 Increase 10/300 GL column, GE Healthcare) pre-equilibrated with the gel filtration buffer (20 mM HEPES-K (pH 7.5), 200 mM KCl, 1 mM TCEP and 5% (v/v) glycerol) containing 1 mM $MgCl_2$. The peak fractions were concentrated by spin concentration (3-kDa cutoff, Merck Millipore), quenched in 1X SDS-Loading dye (2% w/v SDS, 0.1% w/v bromophenol blue and 10% v/v glycerol) and boiled down to 20 µl before loading onto a 4-20% gradient SDS-PAGE.

Mass spectrometry

Protein samples were analyzed using a Synapt mass spectrometer as described elsewhere (53).

RESULTS

A machine learning model for anti-CRISPR prediction

A major challenge in the discovery of new anti-CRISPR proteins is the diversity of amino acid sequences that have been discovered so far, and the lack of predictable structural features between them (54, 55). While some Acrs and *aca* genes are predicted to contain an HTH fold (13, 24, 54, 56, 57), there is no broadly unifying structural motif, making traditional searching methods (such as BLAST similarity searching (50) poorly equipped to identify new Acr families. To address this challenge, we have developed AcRanker, a machine learning model that accepts a proteome as input and ranks its constituent proteins in decreasing order of their expected Acr character.

To build the model, we used EXtreme Gradient Boosting (XGBoost) based ranking (47) with 1-, 2- and 3-mer amino acid composition as input features (43). Other features were considered, but did not improve model performance, or were impractical to include (e.g. requiring experimental data to determine transcription or translation rates). Additionally, the use of sequence features alone can indirectly capture information about the structure of the protein and other properties, such as the isoelectric point and physiochemical properties, while being minimally restrictive. The utility of sequence features has been demonstrated previously (58), including work to predict binding sites within calmodulin (59), where the target proteins sequences are diverse.

To train the model we created a dataset comprised of 20 experimentally verified Acrs taken from the anti-CRISPRdb (37) (Table S1) and their source proteomes. Testing was performed on an additional set of 20 known Acrs, with different predicted mechanisms, sequence composition, and source organisms (Table S2).

Cross-validation by single proteome omission

To evaluate the performance of AcRanker, we performed leave-one-out cross-validation using the training dataset. Out of the 20 known Acr proteomes tested individually, we observed that the ranking-based model ranked seven Acrs higher than other proteins in their respective proteomes (Table 1). In total, 14 out of the 20 known Acrs are ranked within the top 5% in their respective proteomes (Table 1).

Generally, we observe that the machine learning rankings for Acrs contained in phage proteomes are much better than those contained in bacterial proteomes, likely due to their smaller size (Table 1). To test if the relative rankings of the known Acrs found within bacterial proteomes would improve in the context of only prophage-derived proteins, we identified which proteins in the bacterial proteomes were found within prophages using PHASTER (49) and used only that subset to test both models. With the prophage subsets we did observe a higher ranking for the known Acrs due to the removal of higher-ranking proteins not found in the predicted prophages (Table 1).

As a baseline, we also compared the rankings obtained from the machine learning model to a blastp (50) ranking (Table 1). For each excluded Acr in the leave-one-out train/test cycles, the excluded Acrs proteome was used as a query set to BLAST against the 19 other Acrs used for training and the resulting e-values ranked from lowest to highest. These BLASTp scores represent a naïve search strategy that AcRanker seeks to improve upon. The BLAST search method, however, only returned the highest rank for the AcrIF6 family because three distant homologs (using the <40% identity threshold) were included in the training dataset. Interestingly, we also observed that the BLAST method gave higher ranks than AcRanker for AcrIF9, AcrIIA5, and AcrIIA1 (13, 17, 20). However, with the exception of AcrIF6, the BLAST rankings of all the Acrs fell outside of the top 5%, demonstrating the diversity of Acr families, the difficulty of predicting new Acrs *de novo*, and improvement gained using AcRanker.

We next asked which of the features used in AcRanker had the biggest impact on Acr ranking to determine if any biological insight could be gained. Performing a SHAP (SHapley Additive exPlanations) (60) analysis on the constructed model (Figure S1) revealed that the three highest impact features were the presence or absence of three single amino acids: proline, glutamine, and leucine. However, the 'blackbox' nature of machine learning models, the relative continuity of the top 20 impact values, and the lack of a clear relationship between them prevent any clear conclusions from being drawn.

Independent set validation

To validate AcRanker, we used an independent testing dataset of 20 recently discovered Acrs not part of the training dataset (Table S2). Of these 20 Acrs, three are found in phage (AcrIF14, AcrIIA6, and AcrIIIB1) and 10 (AcrIE4-F7, AcrIF11, AcrIF11.1, AcrIF11.2, AcrIC1, AcrIIA3, AcrIIA13, AcrIIC5, AcrVA1, and

AcrVA4) were predicted to be in a prophage region using PHASTER. For the proteins predicted to be in a prophage both the complete bacterial and phage proteome were ranked with AcRanker, otherwise only the complete proteome was ranked (Table S9). The results from the complete bacterial proteomes did generally not perform well (Table S9), with only four (AcrIE5, AcrIC1, AcrIIA3, and AcrIIC5) out of 16 receiving ranks within the top 10. However, of the 13 proteins found within a phage/prophage, AcRanker ranked six within the top five, including two with the highest rank (Table 2).

Within the 20 Acr independent test set, AcRanker returns a higher rank for the majority of (pro-)phage proteomes compared to blastp searching (Table 2). Of the six cases where blastp ranked the known Acr higher than AcRanker, three (AcrIIA6, AcrIIIB1, AcrVA4) were ranked outside of the top 40% by both blastp and AcRanker, and would be unlikely to be discovered using either method. In two of the remaining three cases where blastp returned the higher rank (AcrIE4-F7 and AcrIF11), AcRanker was able to rank at least one member of the family within the top 10 of its respective the predicted prophage proteome. AcrIF14 was the only case where blastp was able to rank the known Acr in the top 10 and AcRanker was not (Table 2). Generally, we observe better performance of AcRanker relative to blastp to identify Acrs, although the appearance of highly ranking known Acrs using blastp suggests a possibility that direct BLAST searching, as opposed to guilt-by-association searching, may be beneficial to locating certain undiscovered Acrs, for which there is some related precedent where three Acr families shared a homologous N-terminus (24).

anti-CRISPR candidate selection

Encouraged by the number of highly ranked Acrs from the test dataset, we proceeded to apply AcRanker to predict novel anti-CRISPRs from self-targeting genomes. Given the ubiquity of *Streptococcus pyogenes* Cas9 (SpyCas9) in gene editing and our inclusion of known SpyCas9 Acrs in the machine learning training dataset (AcrIIA1, AcrIIA2, AcrIIA4, AcrIIA5), we chose to focus specifically on *Streptococcus* species containing Cas9 proteins homologous to SpyCas9.

We began by generating a list of *Streptococcus* genomes containing at least one self-targeting type II-A CRISPR system using Self-Target Spacer Searcher, which has been previously described (11). We found 385 instances of self-targeting from type II-A CRISPR arrays occurring within 241 *Streptococcus* genome assemblies, six of which contained known Acrs. Of these 241 self-targeting arrays, we looked for instances

where the target sequence was flanked by the 3′ NRG protospacer adjacent motif (PAM) characteristic of SpyCas9 and observed that it was present in 20 genomes. These 20 self-targeting arrays would be expected to be lethal for close homologs of SpyCas9, suggesting that other factors, such as the presence of Acrs (11), are preventing CRISPR self-targeting and cell death (Table S10). During our original search of these 20 genomes, *Streptococcus iniae* strain UEL-Si1 was the only one that contained a previously discovered Acr, AcrIIA3 (13), providing a large proteome space to search for novel *acr* genes.

To identify new *acr* gene candidates, we first used PHASTER (49) to predict all of the prophages residing within the 20 self-targeting *Streptococcus* genomes as well as an additional *Listeria monocytogenes* genome (strain R2-502) containing a type II-A self-targeting CRISPR system (with six self-targets) and three well-known AcrIIA genes (13). We included the *Listeria* strain to determine if the known Acrs within it were returned as the top ranked genes, and if not, test the higher-ranking genes as potential additional Acrs within a known Acr-harboring strain. We created lists of the annotated proteins found within each genome's set of prophages. These proteins lists were then ranked with AcRanker to predict the 10 highest ranked genes most likely to be an *acr* (Table S11). Of the approximately 200 genes returned, a subset was selected for further biochemical testing. The selection was based on previous observations that many Acrs are typically short genes with transcripts driven by strong promoters and ribosome binding sites that frequently end with intrinsic terminator sequences (11, 13, 24) (Figure 1). We also looked for proteins encoded in operons with other *acr* or *aca* genes, although this was rare, highlighting a challenge of guilt-by-association approaches.

As with the previous testing dataset, we observed that the known *acr* genes were highly ranked within the test proteomes. Interestingly, a few proteins contained in the same, or overlapping, transcripts as the known Acrs ranked higher with AcRanker (ML1 and ML2). We took these candidates as well as eight others (ML3-ML10) containing the features described above (Figure 1).

Biochemical validation of novel Acrs identified by AcRanker

To determine if the identified proteins were inhibitors of SpyCas9, we purified each candidate and tested their ability to directly inhibit DNA targeting *in vitro*. Of the ten candidate inhibitors, nine were successfully cloned, expressed and purified (Figure S2A and S2B). To assess inhibition of DNA targeting *in vitro*, we

first assayed the ability of SpyCas9 to cleave double stranded DNA (dsDNA) when incubated in the presence of a 50-fold excess of each candidate Acr (Figure 2A). While SpyCas9 was capable of complete DNA target cleavage, the generation of DNA cleavage products was attenuated in the presence of the positive control inhibitor AcrIIA4 and the candidates ML1 or ML8. To determine the potency of inhibition, we tested the ability of SpyCas9 to cleave the DNA target in the presence of a dilution series of ML1 or ML8 (Figure 2B). In contrast to AcrIIA4, an established potent inhibitor of SpyCas9 (13), both ML1 and ML8 inhibited SpyCas9 with around a 10-fold lower potency. We wondered if the high concentration of ML1 or ML8 required to completely inhibit Cas9 might represent an *in vitro* concentration-dependent artifact. To explore this, we assayed SpyCas9 DNA cleavage against a titration series of either non-target DNA competitor, BSA, ML2, or ML3 and observed no significant inhibition of SpyCas9, even with a 100-fold excess (Figure S3B-D). Taken together, these data indicated that both ML1 and ML8 weakly inhibit SpyCas9 DNA cleavage *in vitro*.

We next tested the ability of the AcRanker-generated candidates to inhibit *Staphylococcus aureus* (SauCas9), another Cas9 commonly used for gene editing (61, 62) to determine whether any of the candidates identified from self-targeting *Streptococcus* genomes had broader Cas9 inhibition activity. At a 25-fold excess relative to the SauCas9 RNP complex, ML3 and ML8 were able to inhibit SauCas9 dsDNA cleavage (Figure 2C). To determine potency, we incubated a dilution series of either ML3 or ML8 with SauCas9 before the addition of the DNA target. However, in comparison to AcrIIA5, an established strong inhibitor of SauCas9 (20, 24, 63), both Acr candidates inhibited SauCas9 with approximately 50-fold lower potency (Figure 2D, Figure S4A, S4B), an activity we confirmed was not due to a false positive from the high concentration of protein in the assay (Figure S4A).

Given the relatively weak inhibition of both SpyCas9 and SauCas9, we next tested the specificity of ML1, ML3, and ML8 by assaying their ability to block DNA targeting by either AsCas12a or the restriction enzyme AlwNI. Neither AcrIIA4, ML1, ML3, nor ML8 were able to inhibit DNA targeting by AlwNI, suggesting that they all are specific inhibitors of CRISPR effectors (Figures S5A and S5B). Consistent with this, inhibition of AsCas12a was only observed with ML1 and ML8 at a 100-fold excess (Figure S5C). Taken together, our data show that ML1, ML3, and ML8 are low potency inhibitors of SpyCas9 (ML1 and ML8) or SauCas9

(ML3 and ML8). While testing ML1-ML10 for Acr activity, Osuna, et al. described AcrIIA12, a specific inhibitor of LmoCas9 in plaque assays, which shares the same sequence as ML3 (25).

ML1: a potent inhibitor of SinCas9

ML1 was identified in the *Streptococcus iniae* (Sin) genome. Previous studies have reported anti-CRISPRs can exhibit either selective or broad-spectrum inhibition of divergent Cas effectors (14, 33). Given that SinCas9 is ~70% identical to SpyCas9 and only ~26% identical to SauCas9 we wondered whether ML1 is a more potent inhibitor of SinCas9. To explore this, we cloned, expressed, and purified SinCas9 protein for use in *in vitro* DNA targeting assays. Like SpyCas9, SinCas9 was capable of cleaving dsDNA targets proximal to an NGG PAM using a sgRNA derived from a fusion of the tracrRNA and crRNA (Figure 3A, Figure S6, Figure S7). Similar to SpyCas9, both ML1 and ML8 inhibited DNA cleavage by SinCas9 (Figure 3A). Using a titration of ML1 and ML8, we again assayed the potency of SinCas9 inhibition (Figure 3B, Figure S6B). Strikingly, in contrast to the weak inhibition of SpyCas9, ML1 was able to potently inhibit DNA cleavage by SinCas9 (Figure 3B). To investigate at which step ML1 inactivates SinCas9 function, we carried out *in vitro* cleavage assays where ML1 was incubated with SinCas9 before and after the addition of sgRNA (Figure S6C). In both cases the DNA cleavage activity of SinCas9 was potently inhibited, suggesting that ML1 inhibits activity after sgRNA binding to Cas9.

A number of reported type-IIA Acrs inhibit their cognate Cas9 by competing with target DNA through PAM mimicry (52, 64). We noted that SinCas9 was susceptible to inhibition by AcrIIA4 at 100-fold excess (Figure 3A) and AcrIIA2 at 10-fold excess (Figure S6D), both PAM mimics that inhibit PAM recognition by SpyCas9 (15, 52). Like these established PAM mimics, ML1 is a small protein with a predicted negatively charged surface potential (isoelectric point of 4.3), suggesting that it too might compete with target DNA. To explore this idea, we developed a competition binding experiment to assay if the association of ML1 with SinCas9 might prevent the binding of AcrIIA2 (Figure 4A). First, we incubated either AcrIIA2 or ML1 with the SinCas9-sgRNA complex and observed a stable SinCas9-sgRNA-Acr complex on a gel filtration column (Figure 4B, Figure S8A) with the complex components all resolvable on a protein gel (Figure 4C, Figure S8B). To determine if ML1 binding to the SinCas9 RNP could prevent AcrIIA2 binding, we first formed the SinCas9-sgRNA-ML1 complex and then incubated with AcrIIA2 before resolving over a column.

Incubating ML1 with the SinCas9 RNP before adding AcrIIA2 abolished AcrIIA2 co-elution with SinCas9-sgRNA (Figure 4C, Figure S8B), suggesting that ML1 might occupy the same site on SinCas9. Collectively, these data are consistent with a model where ML1 directly binds to the SinCas9-sgRNA complex to form a complex that is incompatible with AcrIIA2's ability to bind to the PAM interacting domain (52).

DISCUSSION

With the growth of the anti-CRISPR field, there has been a need for improved tools to search the extensive proteomic space to find new anti-CRISPRs more efficiently. In this work we developed a machine learning method, AcRanker, as a first step toward the direct prediction of *acr* genes *de novo* with minimal knowledge *a priori.* We show that with only protein sequence features, AcRanker is able to highly rank Acrs from within prophage proteomes. Using a combination of AcRanker and self-targeting information from STSS (11), we were able to quickly reduce to a few top *acr* gene candidates for direct synthesis and testing of anti-CRISPR properties. From these candidates, we identified two novel Acrs: here named AcrIIA20 and AcrIIA21. AcrIIA20 (ML1) inhibits *Streptococcus iniae* Cas9 (SinCas9) with high potency and *Streptococcus pyogenes* Cas9 (SpyCas9) with low potency. With only 64 amino acids and a molecular weight of 7.3 kDa, to our knowledge it is the smallest type II Acr found to date. Based on the negative charge of AcrIIA20 and its competitive binding with AcrIIA2, we speculate that AcrIIA20 inhibits Cas9 dsDNA cleavage via a similar mechanism of PAM mimicry. In addition, we found AcrIIA21 (ML8), a broadly acting type II-A Acr, which is able to inhibit SpyCas9, SauCas9 as well as SinCas9, although with low potency.

The narrow and broader inhibition range of AcrIIA20 and AcrIIA21, respectively, is mirrored in their distribution in other genomes. Within the NCBI protein database, only a handful of homologs can be found for AcrIIA20 in closely related *Streptococcus* species (namely *iniae, uberis,* and *dysgalactiae*). In contrast, sequences sharing homology with AcrIIA21 are found broadly in *Lactobacillales* and beyond, owing at least in part to its shared identity with replication initiator protein A, a single stranded DNA binding protein, suggesting nucleic acid binding as one potential mechanism of inhibition for AcrIIA21.

We also observe weak inhibition of SauCas9 with ML3 (AcrIIA12), which was shown to be a specific inhibitor of *Listeria monocytogenes* Cas9 (LmoCas9) while this study was being conducted (25). Because we were unable to test LmoCas9 (due to the difficulty of purifying it intact and active), we were unable to observe strong inhibition activity specific to its host Cas9. Similarly, we were unable to satisfactorily purify *S. agalactiae* Cas9 (SagCas9) to test ML4-ML10 against the Cas9 found in the same genomes in which they were found, leaving the door open for the possibility that they are specific against SagCas9.

AcRanker adds yet another tool to the anti-CRISPR hunter's toolbox by providing an alternative to BLAST and guilt-by-association searching to find new Acr families. In fact, we find that of the three

candidates that we or others validated (ML1, ML3, and ML8), all had significantly higher rankings with AcRanker over BLAST (Table S12). However, we do see some cases where BLAST ranks known Acrs higher than AcRanker (Tables 1 and 2), providing a potential complementary approach, although one we believe is less likely to lead to new Acrs.

The ability to identify potential new Acr candidates directly from protein sequence with AcRanker opens the door for testing many new proteins without the need for laborious screening efforts. Searching within prophages of genomes containing self-targeting CRISPR arrays promises to be particularly effective, as the potential inhibitors for a specific CRISPR system can be quickly ranked to make a short list of candidates to test. We expect that direct Acr prediction methods like AcRanker will continue to reveal many more Acrs distributed across many bacterial species, finding new Acrs with unique properties for yet unforeseen future biotechnology applications.

DATA AVAILABILITY

A webserver implementation of AcRanker is publicly available at http://acranker.pythonanywhere.com/. The Python code for the webserver implementation is available in the GitHub repository (https://github.com/amina01/AcRanker).

SUPPLEMENTARY DATA

Supplementary data are available at NAR online.

ACKNOWLEDGEMENTS

We thank Blake McMahon for plasmid cloning and protein purification. We thank Haridha Shivram and Patrick Pausch for providing useful tips throughout the project. We also want to thank Dylan Smock for expressing proteins and Brittney W. Thornton for technical advice.

FUNDING

The authors acknowledge financial support from the Defense Advanced Research Projects Agency (DARPA) (award HR0011-17-2-0043 to J.A.D.), the Paul G. Allen Frontiers Group and the National Science Foundation (MCB-1244557 to J.A.D.). J.A.D. is an investigator of the Howard Hughes Medical Institute (HHMI), and this study was supported in part by HHMI; J.A.D is also a Paul Allen Distinguished Investigator. A mass spectrometer was purchased using National Institutes of Health support (grant number 1S10OD020062-01). Amina Asif is funded via Information Technology and Telecommunication Endowment Fund at Pakistan Institute of Engineering and Applied Sciences.

CONFLICT OF INTEREST

J.A.D. is a co-founder of Caribou Biosciences, Editas Medicine, Intellia Therapeutics, Scribe Therapeutics, and Mammoth Biosciences, a scientific adviser to Caribou Biosciences, Intellia Therapeutics, Scribe Therapeutics, Synthego, Felix Biosciences, Inari, Mammoth Biosciences, and eFFECTOR Therapeutics, and a director of Johnson & Johnson and has sponsored research projects supported by Pfizer and Biogen. The Regents of the University of California have patents pending for CRISPR related technologies on which the authors are inventors.

AUTHOR CONTRIBUTIONS

Conceptualization, F.A.A.M., K.E.W., J.A.D.; Methodology, A.A., K.E.W., S.E., G.J.K., F.A.A.M.; Software, A.A., F.A.A.M., K.E.W.; Investigation, A.A., K.E.W., S.E., F.A.A.; Biochemical Analysis, S.E., G.J.K., A.T.I.; Data Curation, A.A., K.E.W., S.E.; Writing, A.A., S.E., K.E.W., G.J.K., F.A.A.M.; Funding Acquisition, K.E.W., S.E., G.J.K., J.A.D., F.A.A.M., A.T.I.

Table 1. Results for leave-one-out cross-validation. Each row of the table indicates which Acr was excluded from the training dataset and used as a test dataset, and each number displayed is the ranking of the known Acr received from the indicated test proteome using either the blastp search against all other known Acrs (BLAST) or AcRanker. The Acrs from bacterial proteomes - AcrIF6, AcrIF9, AcrIF10, AcrIIA1, AcrIIA2, and AcrIIA4 - were also ranked using only the subset of proteins predicted to reside within prophages as predicted by PHASTER (49). Two Acrs from bacterial proteomes did not occur in the predicted prophages (WP_014702809.1 and WP_031500045.1) and are indicated by dash placeholders. All three prophage proteome subset fields have been left empty for Acrs from phage proteomes.

		Complete Proteome			Prophage Subset		
Accession No.	**Anti-CRISPR family**	**Proteome Size**	**BLAST rank**	**AcRanker rank**	**Proteome Size**	**BLAST rank**	**AcRanker rank**
YP_007392738.1	AcrIE1	57	33	**1**	-	-	-
YP_007392439.1	AcrIE2	54	18	2	-	-	-
YP_950454.1	AcrIE3	52	17	**1**	-	-	-
NP_938238.1	AcrIE4	54	11	**1**	-	-	-
YP_007392342.1	AcrIF1	56	21	11	-	-	-
YP_002332454.1	AcrIF2	51	34	**1**	-	-	-
YP_007392440.1	AcrIF3	54	5	**1**	-	-	-
YP_007392799.1	AcrIF4	57	36	3	-	-	-
YP_007392740.1	AcrIF5	57	26	19	-	-	-
WP_043884810.1	AcrIF6	6095	**1**	80	361	**1**	15
WP_019933870.1	AcrIF6	3045	**1**	13	72	**1**	**1**
WP_014702809.1	AcrIF6	2689	**1**	130	57	-	-
ACD38920.1	AcrIF7	57	20	**1**	-	-	-
AFC22483.1	AcrIF8	68	30	**1**	-	-	-
WP_031500045.1	AcrIF9	4928	198	333	37	-	-
KEK29119.1	AcrIF10	3552	189	17	70	23	2
AEO04364.1	AcrIIA1	2951	183	770	146	60	87
AEO04363.1	AcrIIA2	2952	210	16	146	34	3
AEO04689.1	AcrIIA4	2951	59	21	146	9	4
ASD50988.1	AcrIIA5	54	5	8	-	-	-

Table 2. Independent testing set validation results. Thirteen proteomes containing non-redundant (<40% sequence identity) Acrs from phage or bacterial prophage (as predicted by PHASTER) were ranked with either AcRanker or a blastp search against the training set of Acrs.

		Prophage Subset		
Accession no.	**Anti-CRISPR family**	**Proteome Size**	**BLAST rank**	**AcRanker rank**
WP_064584002.1	AcrIE4-F7	111	**1**	4
WP_038819808.1	AcrIF11	64	38	3
WP_033936089.1	AcrIF11.1	92	38	**1**
EGE18857.1	AcrIF11.2	59	**1**	30
AKI27193.1	AcrIF14	68	5	14
WP_046701304.1	AcrIC1	72	15	**1**
WP_014930691.1	AcrIIA3	74	10	2
WP_149028791.1	AcrIIA6	40	21	23
AKS70260.1	AcrIIA13	145	29	3
WP_002642161.1	AcrIIC5	367	237	6
NP_666582.1	AcrIIIB1	54	25	44
WP_046701302.1	AcrVA1	72	18	10
WP_046699156.1	AcrVA4	293	181	220

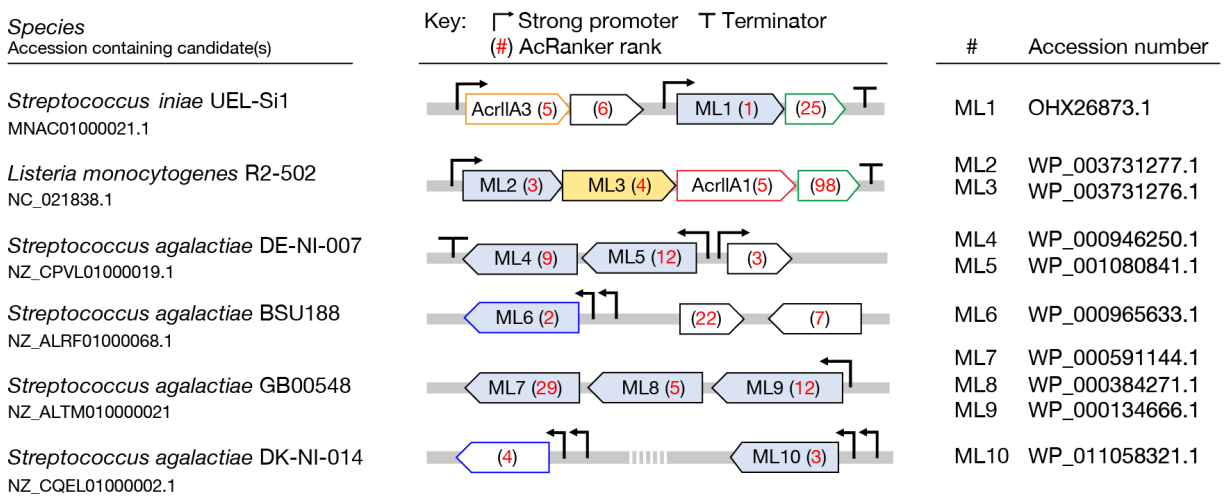


Figure 1. Acr candidates selected for biochemical testing. Ten Acr candidates were selected from manual inspection for further biochemical testing (blue fill). Each candidate is shown in its genomic context with its assigned rank from AcRanker noted in red. Homologous proteins share the same color border (green, blue). Homologs of AcrIIA3 (orange border) and AcrIIA1 (red border) are indicated. While testing the ML candidates, ML3 (yellow fill) was identified as a specific inhibitor of LmoCas9 (25).

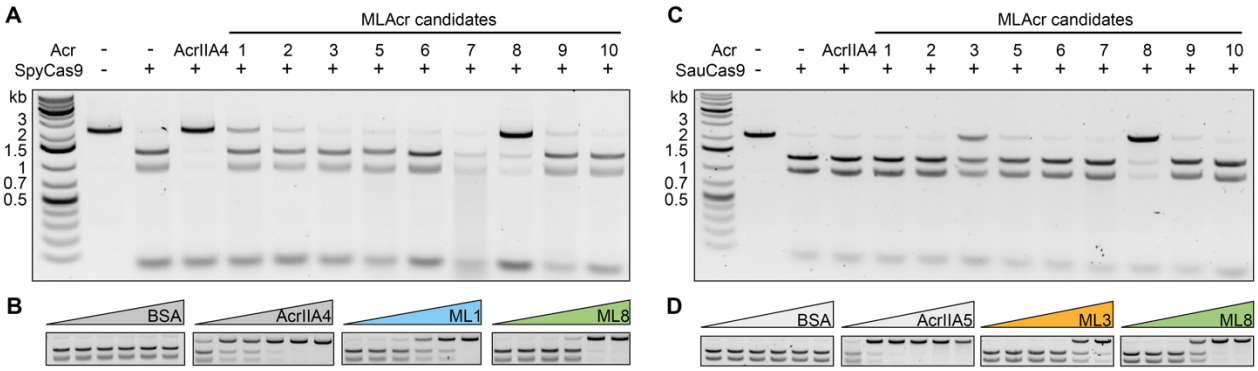
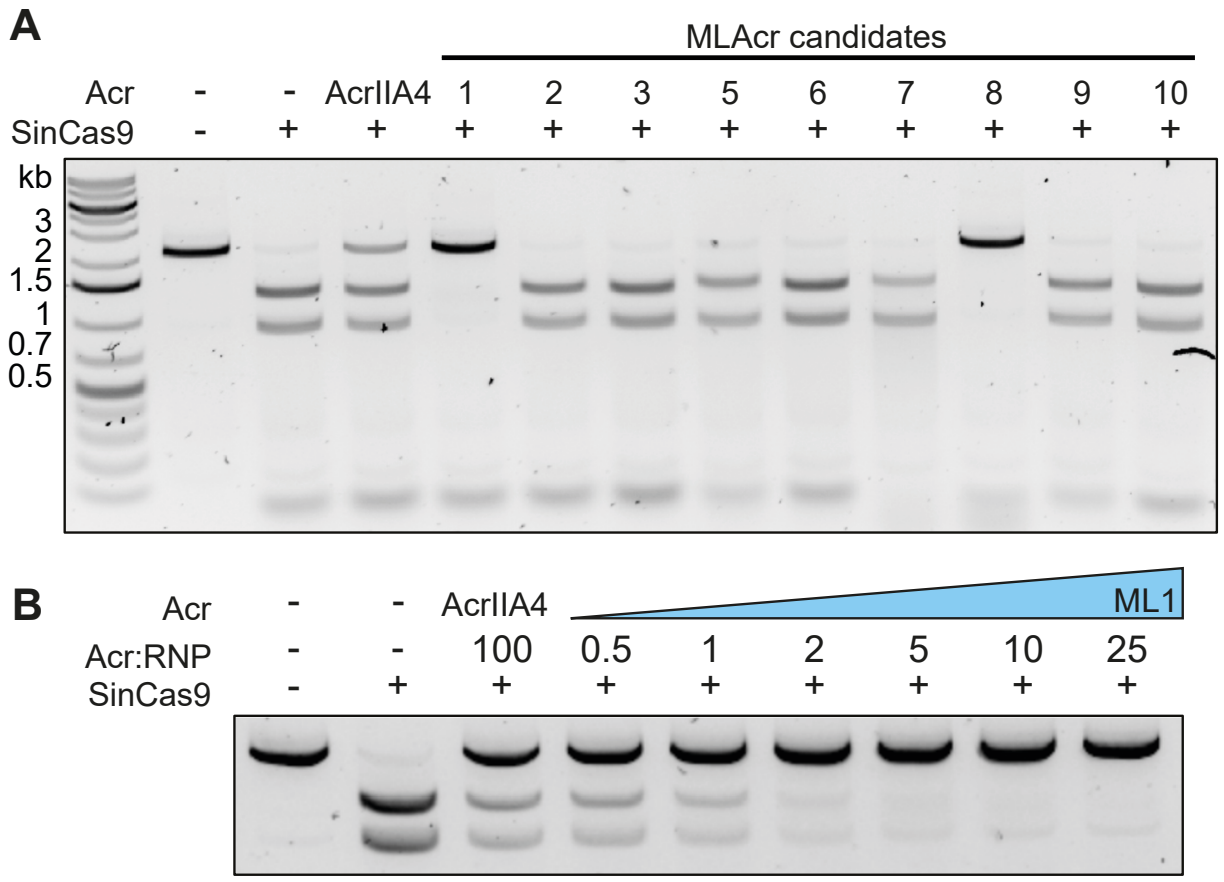


Figure 2. Inhibition of SpyCas9 and SauCas9 by newly discovered Acr candidates. (A) *In vitro* cleavage of dsDNA by SpyCas9 in the absence or presence of a 50-fold excess of AcrIIA4 (positive control) and each Acr candidate. (**B**) *In vitro* cleavage of dsDNA by SpyCas9 in the presence of increasing concentrations of (left to right) BSA (negative control), AcrIIA4 (positive control), ML1 and ML8 (Acr:RNP 0.1-, 1-, 2- ,10-, 50- and 100-fold excess from left to right). (**C**) *In vitro* cleavage of dsDNA by SauCas9 in the absence or presence of a 25-fold excess of each Acr candidate. (**D**) *In vitro* cleavage of dsDNA by SauCas9 in the presence of increasing concentrations of (left to right) BSA (negative control), AcrIIA5 (positive control, Acr:RNP 0.1-, 1-, 2- ,4-, 8- and 10-fold excess from left to right), ML3 and ML8 (Acr:RNP 0.1-, 1-, 2- ,10-, 50- and 100-fold excess from left to right). Uncropped gel images for panels B and D are shown in Figures S3 and S4.



507

508 **Figure 3. ML1 and ML8 inhibit SinCas9 with ML1 showing high potency. (A)** *In vitro* cleavage of dsDNA
509 by SinCas9 in the absence or presence of a 50-fold excess of each Acr candidate. (**B**) *In vitro* cleavage of
510 dsDNA by SinCas9 in the presence of increasing concentrations of ML1. The uncropped gel image for
511 panel B is shown in Figure S6.

512

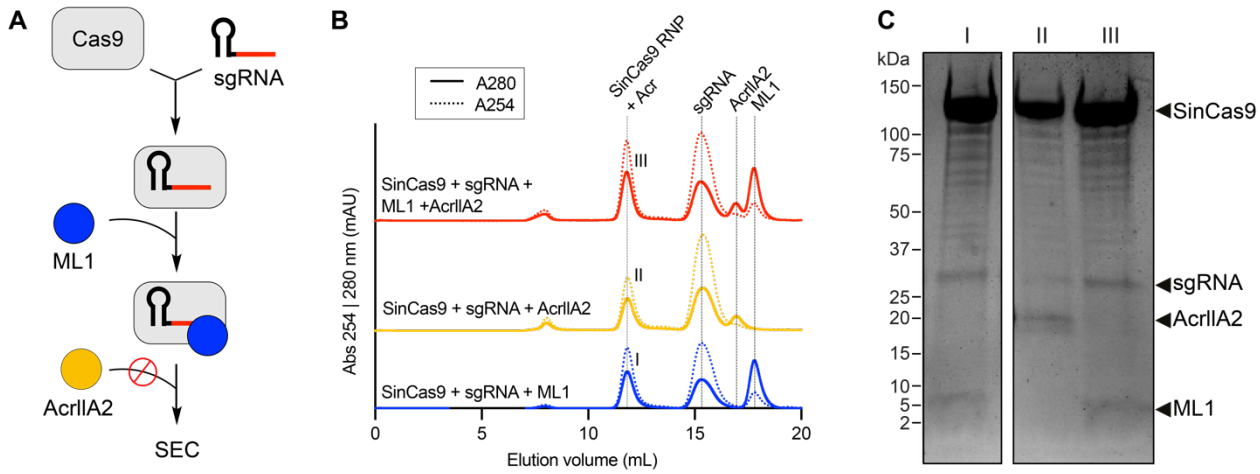


Figure 4. ML1 competes with AcrIIA2 to bind to the SinCas9-sgRNA complex. (A) Flowchart for the competition binding experiment between ML1 and AcrIIA2. Binding of the Acr to the SinCas9-sgRNA RNP was reconstituted using size-exclusion chromatography (SEC). (**B**) Size-exclusion chromatogram of SinCas9-sgRNA in the presence of either ML1, AcrIIA2 or both Acrs with AcrIIA2 added after ML1. (**C**) Coomassie-stained polyacrylamide gel illustrating the components of the SinCas9-RNP fraction annotated (I), (II), and (III) in panel B.

REFERENCES

1. Bolotin,A., Quinquis,B., Sorokin,A. and Dusko Ehrlich,S. (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*, **151**, 2551–2561.
2. Horvath,P. and Barrangou,R. (2010) CRISPR/Cas, the immune system of bacteria and archaea. *Science*, **327**, 167–170.
3. Barrangou,R. (2015) The roles of CRISPR-Cas systems in adaptive immunity and beyond. *Curr. Opin. Immunol.*, **32**, 36–41.
4. Knott,G.J. and Doudna,J.A. (2018) CRISPR-Cas guides the future of genetic engineering. *Science*, **361**, 866–869.
5. Song,G., Jia,M., Chen,K., Kong,X., Khattak,B., Xie,C., Li,A. and Mao,L. (2016) CRISPR/Cas9: A powerful tool for crop genome editing. *Crop Journal*, **4**, 75–82.
6. Ledford,H. (2016) CRISPR: gene editing is just the beginning. *Nature*, **531**, 156–159.
7. Zhang,F., Wen,Y. and Guo,X. (2014) CRISPR/Cas9 for genome editing: progress, implications and challenges. *Hum. Mol. Genet.*, **23**, R40–R46.
8. van Diemen,F.R., Kruse,E.M., Hooykaas,M.J.G., Bruggeling,C.E., Schürch,A.C., van Ham,P.M., Imhof,S.M., Nijhuis,M., Wiertz,E.J.H.J. and Lebbink,R.J. (2016) CRISPR/Cas9-Mediated Genome Editing of Herpesviruses Limits Productive and Latent Infections. *PLoS Pathogens*, **12**, e1005701.
9. Doudna,J.A. and Charpentier,E. (2014) The new frontier of genome engineering with CRISPR-Cas9. *Science*, **346**, 1258096.
10. Bondy-Denomy,J., Pawluk,A., Maxwell,K.L. and Davidson,A.R. (2013) Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature*, **493**, 429–432.
11. Kyle E. Watters, Christof Fellmann, Hua B. Bai, Shawn M. Ren and Jennifer A. Doudna (2018) Systematic discovery of natural CRISPR-Cas12a inhibitors. *Science*, **362**, 236–239.
12. Marino,N.D., Zhang,J.Y., Borges,A.L., Sousa,A.A., Leon,L.M., Rauch,B.J., Walton,R.T., Berry,J.D., Joung,J.K., Kleinstiver,B.P., *et al.* (2018) Discovery of widespread type I and type V CRISPR-Cas inhibitors. *Science*, **362**, 240–242.
13. Rauch,B.J., Silvis,M.R., Hultquist,J.F., Waters,C.S., McGregor,M.J., Krogan,N.J. and Bondy-Denomy,J. (2017) Inhibition of CRISPR-Cas9 with Bacteriophage Proteins. *Cell*, **168**, 150–158.
14. Harrington,L.B., Doxzen,K.W., Ma,E., Liu,J.J., Knott,G.J., Edraki,A., Garcia,B., Amrani,N., Chen,J.S., Cofsky,J.C., *et al.* (2017) A Broad-Spectrum Inhibitor of CRISPR-Cas9. *Cell*, **170**, 1224–1233.
15. Shin,J., Jiang,F., Liu,J.-J., Bray,N.L., Rauch,B.J., Baik,S.H., Nogales,E., Bondy-Denomy,J., Corn,J.E. and Doudna,J.A. (2017) Disabling Cas9 by an anti-CRISPR DNA mimic. *Sci. Adv.*, **3**, e1701620.
16. Maxwell,K.L. (2017) The Anti-CRISPR Story: A Battle for Survival. *Molecular Cell*, **68**, 8–14.
17. Pawluk,A., Staals,R.H.J., Taylor,C., Watson,B.N.J., Saha,S., Fineran,P.C., Maxwell,K.L. and Davidson,A.R. (2016) Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species. *Nat. Microbiol.*, **1**, 16085.

18. Borges,A.L., Davidson,A.R. and Bondy-Denomy,J. (2017) The Discovery, Mechanisms, and Evolutionary Impact of Anti-CRISPRs. *Annu Rev Virol*, **4**, 37–59.
19. He,F., Bhoobalan-Chitty,Y., Van,L.B., Kjeldsen,A.L., Dedola,M., Makarova,K.S., Koonin,E. V, Brodersen,D.E. and Peng,X. (2018) Anti-CRISPR proteins encoded by archaeal lytic viruses inhibit subtype ID immunity. *Nat. Microbiol.*, **3**, 461–469.
20. Hynes,A.P., Rousseau,G.M., Lemay,M.L., Horvath,P., Romero,D.A., Fremaux,C. and Moineau,S. (2017) An anti-CRISPR from a virulent streptococcal phage inhibits Streptococcus pyogenes Cas9. *Nat. Microbiol.*, **2**, 1374–1380.
21. Hynes,A.P., Rousseau,G.M., Agudelo,D., Goulet,A., Amigues,B., Loehr,J., Romero,D.A., Fremaux,C., Horvath,P., Doyon,Y., *et al.* (2018) Widespread anti-CRISPR proteins in virulent bacteriophages inhibit a range of Cas9 proteins. *Nat. Commun.*, **9**, 2919.
22. Pawluk,A., Bondy-Denomy,J., Cheung,V.H.W., Maxwell,K.L. and Davidson,A.R. (2014) A New Group of Phage Anti-CRISPR Genes Inhibits the Type I-E CRISPR-Cas System of Pseudomonas aeruginosa. *mBio*, **5**, e00896-14.
23. Pawluk,A., Shah,M., Mejdani,M., Calmettes,C., Moraes,T.F., Davidson,A.R. and Maxwell,K.L. (2017) Disabling a Type I-E CRISPR-Cas Nuclease with a Bacteriophage-Encoded Anti-CRISPR Protein. *MBio*, **8**, e01751-17.
24. Watters,K.E., Shivram,H., Fellmann,C., Lew,R.J., McMahon,B. and Doudna,J.A. (2020) Potent CRISPR-Cas9 inhibitors from Staphylococcus genomes. *Proc. Natl. Acad. Sci. USA*, **117**, 1–9.
25. Osuna,B.A., Karambelkar,S., Mahendra,C., Christie,K.A. and others (2019) Listeria phages induce Cas9 degradation to protect lysogenic genomes. *bioRxiv*.
26. Uribe,R. V, van der Helm,E., Misiakou,M.A., Lee,S.W., Kol,S. and Sommer,M.O.A. (2019) Discovery and Characterization of Cas9 Inhibitors Disseminated across Seven Bacterial Phyla. *Cell Host and Microbe*, **25**, 233–241.
27. Forsberg,K.J., Bhatt,I. V, Schmidtke,D.T., Javanmardi,K., Dillard,K.E., Stoddard,B.L., Finkelstein,I.J., Kaiser,B.K. and Malik,H.S. (2019) Functional metagenomics-guided discovery of potent Cas9 inhibitors in the human microbiome. *Elife*, **8**, e46540.
28. Lee,J., Mir,A., Edraki,A., Garcia,B., Amrani,N., Lou,H.E., Gainetdinov,I., Pawluk,A., Ibraheim,R., Gao,X.D., *et al.* (2018) Potent Cas9 inhibition in bacterial and human cells by AcrIIC4 and AcrIIC5 anti-CRISPR proteins. *mBio*, **9**, 1–17.
29. Heussler,G.E. and O'Toole,G.A. (2016) Friendly fire: Biological functions and consequences of chromosomal targeting by CRISPR-cas systems. *Journal of Bacteriology*, **198**, 1481–1486.
30. Pawluk,A., Davidson,A.R. and Maxwell,K.L. (2018) Anti-CRISPR: discovery, mechanism and function. *Nat. Rev. Microbiol.*, **16**, 12–17.
31. Bondy-Denomy,J., Garcia,B., Strum,S., Du,M., Rollins,M.F., Hidalgo-Reyes,Y., Wiedenheft,B., Maxwell,K.L. and Davidson,A.R. (2015) Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins. *Nature*, **526**, 136–139.

32. Maxwell,K.L. (2016) Phages Fight Back: Inactivation of the CRISPR-Cas Bacterial Immune System by Anti-CRISPR Proteins. *PLoS Pathog.*, **12**, e1005282.
33. Knott,G.J., Thornton,B.W., Lobba,M.J., Liu,J., Al-Shayeb,B., Watters,K.E. and Doudna,J.A. (2019) Broad-spectrum enzymatic inhibition of CRISPR-Cas12a. *Nat. Struct. Mol. Biol.*, **26**, 315–321.
34. Dong,L., Guan,X., Li,N., Zhang,F., Zhu,Y., Ren,K., Yu,L., Zhou,F., Han,Z., Gao,N., *et al.* (2019) An anti-CRISPR protein disables type V Cas12a by acetylation. *Nat. Struct. Mol. Biol.*, **26**, 308–314.
35. Zhang,H., Li,Z., Daczkowski,C.M., Gabel,C., Mesecar,A.D. and Chang,L. (2019) Structural Basis for the Inhibition of CRISPR-Cas12a by Anti-CRISPR Proteins. *Cell Host and Microbe*, **25**, 815–826.
36. Knott,G.J., Cress,B.F., Liu,J.-J., Thornton,B.W., Lew,R.J., Al-Shayeb,B., Rosenberg,D.J., Hammel,M., Adler,B.A., Lobba,M.J., *et al.* (2019) Structural basis for AcrVA4 inhibition of specific CRISPR-Cas12a. *eLife*, **8**, e49110.
37. Dong,C., Hao,G.F., Hua,H.L., Liu,S., Labena,A.A., Chai,G., Huang,J., Rao,N. and Guo,F.B. (2018) Anti-CRISPRdb: A comprehensive online resource for anti-CRISPR proteins. *Nucleic Acids Res.*, **46**, D393–D398.
38. Walsh,I., Pollastri,G. and Tosatto,S.C.E. (2016) Correct machine learning on protein sequences: A peer-reviewing perspective. *Briefings in Bioinformatics*, **17**, 831–840.
39. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Engineering*, **12**, 85–94.
40. Huang,Y., Niu,B., Gao,Y., Fu,L. and Li,W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.
41. Bhoobalan-Chitty,Y., Johansen,T.B., Di Cianni,N. and Peng,X. (2019) Inhibition of Type III CRISPR-Cas Immunity by an Archaeal Virus-Encoded Anti-CRISPR Protein. *Cell*, **179**, 448-458.e11.
42. Hwang,S. and Maxwell,K.L. (2019) Meet the Anti-CRISPRs: Widespread Protein Inhibitors of CRISPR-Cas Systems. *The CRISPR Journal*, **2**, 23–30.
43. Saidi,R., Maddouri,M. and Mephu Nguifo,E. (2010) Protein sequences classification by means of feature extraction with substitution matrices. *BMC Bioinformatics*, **11**, 175.
44. Shen,J., Zhang,J., Luo,X., Zhu,W., Yu,K., Chen,K., Li,Y. and Jiang,H. (2007) Predicting protein-protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. USA*, **104**, 4337–4341.
45. Leslie,C., Eskin,E. and Noble,W.S. (2002) The spectrum kernel: A string kernel for SVM protein classification. *Proc. of the Pacific Symposium on Biocomputing*.
46. Ben-Hur,A. and Weston,J. (2010) A user's guide to support vector machines. *Methods Mol. Biol.*, **609**, 223–239.
47. Chen,T. and Guestrin,C. (2016) XGBoost: A Scalable Tree Boosting System. *Proc. of the 22nd ACM SIGKDD Int. Conf.*
48. Koonin,E. V. and Makarova,K.S. (2018) Anti-CRISPRs on the march. *Science*, **362**, 156–157.
49. Arndt,D., Grant,J.R., Marcu,A., Sajed,T., Pon,A., Liang,Y. and Wishart,D.S. (2016) PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.*, **44**, W16–W21.

50. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.
51. East-Seletsky,A., O'Connell,M.R., Knight,S.C., Burstein,D., Cate,J.H.D., Tjian,R. and Doudna,J.A. (2016) Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature*, **538**, 270–273.
52. Jiang,F., Liu,J.J., Osuna,B.A., Xu,M., Berry,J.D., Rauch,B.J., Nogales,E., Bondy-Denomy,J. and Doudna,J.A. (2019) Temperature-Responsive Competitive Inhibition of CRISPR-Cas9. *Molecular Cell*, **73**, 601–610.
53. Light,S.H., Su,L., Rivera-Lugo,R., Cornejo,J.A., Louie,A., Iavarone,A.T., Ajo-Franklin,C.M. and Portnoy,D.A. (2018) A flavin-based extracellular electron transfer mechanism in diverse Gram-positive bacteria. *Nature*, **562**, 140–144.
54. Zhang,F., Song,G. and Tian,Y. (2019) Anti-CRISPRs: The natural inhibitors for CRISPR-Cas systems. *Anim Models Exp Med.*, **2**, 69–75.
55. Bondy-Denomy,J., Davidson,A.R., Doudna,J.A., Peter C.fineran, Maxwell,K.L., Moineau,S., Peng,X., Sontheimer,E.J. and Wiedenheft,B. (2018) A Unified Resource for Tracking Anti-CRISPR Names. *The CRISPR Journal*, **1**, 304–305.
56. Ka,D., An,S.Y., Suh,J.Y. and Bae,E. (2018) Crystal structure of an anti-CRISPR protein, AcrIIA1. *Nucleic Acids Res.*, **46**, 485–492.
57. Zhu,Y., Zhang,F. and Huang,Z. (2018) Structural insights into the inactivation of CRISPR-Cas systems by diverse anti-CRISPR proteins. *BMC Biol.*, **16**, 32.
58. Al-Shahib,A., Breitling,R. and Gilbert,D.R. (2007) Predicting protein function by machine learning on amino acid sequences - a critical evaluation. *BMC Genomics*, **8**, 78.
59. Minhas,F.U.A.A. and Ben-Hur,A. (2012) Multiple instance learning of Calmodulin binding sites. *Bioinformatics*, **28**, 416–422.
60. Lundberg,S.M. and Lee,S.-I. (2017) A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*.
61. Ran,F.A., Cong,L., Yan,W.X., Scott,D.A., Gootenberg,J.S., Kriz,A.J., Zetsche,B., Shalem,O., Wu,X., Makarova,K.S., *et al.* (2015) In vivo genome editing using Staphylococcus aureus Cas9. *Nature*, **520**, 186–191.
62. Yourik,P., Fuchs,R.T., Mabuchi,M., Curcuru,J.L. and Robb,G.B. (2019) Staphylococcus aureus Cas9 is a multiple-turnover enzyme. *RNA*, **25**, 35–44.
63. Garcia,B., Lee,J., Edraki,A., Hidalgo-Reyes,Y., Erwood,S., Mir,A., Trost,C.N., Seroussi,U., Stanley,S.Y., Cohn,R.D., *et al.* (2019) Anti-CRISPR AcrIIA5 Potently Inhibits All Cas9 Homologs Used for Genome Editing. *Cell Reports*, **29**, 1739–1746.
64. Yang,H. and Patel,D.J. (2017) Inhibition Mechanism of an Anti-CRISPR Suppressor AcrIIA4 Targeting SpyCas9. *Mol. Cell*, **67**, 117–127.

Supplementary Information

Machine Learning Predicts New Anti-CRISPR Proteins

Simon Eitzinger[1†], Amina Asif[2†], Kyle E. Watters[1†], Anthony T. Iavarone[3], Gavin J. Knott[1], Jennifer A. Doudna[1,4-8*], and Fayyaz ul Amir Afsar Minhas[2,9*]

[1]Department of Molecular and Cell Biology, University of California Berkeley, Berkeley, CA 94720, USA

[2]Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences (PIEAS), PO Nilore, Islamabad, Pakistan.

[3]QB3/Chemistry Mass Spectrometry Facility, University of California, Berkeley, Berkeley, CA, USA

[4]Department of Chemistry, University of California Berkeley, Berkeley, CA, 94720, USA

[5]Innovative Genomics Initiative, University of California Berkeley, Berkeley, CA, 94720, USA

[6]Center for RNA Systems Biology, University of California Berkeley, Berkeley, CA, 94720, USA

[7]Howard Hughes Medical Institute, University of California Berkeley, Berkeley, CA, 94720, USA

[8]Molecular Biophysics & Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA

[9]Department of Computer Science, University of Warwick, Coventry, CV4 7AL, UK

[†] These authors contributed equally to this work.

*To whom correspondence should be addressed: doudna@berkeley.edu, fayyaz.minhas14@alumni.colostate.edu

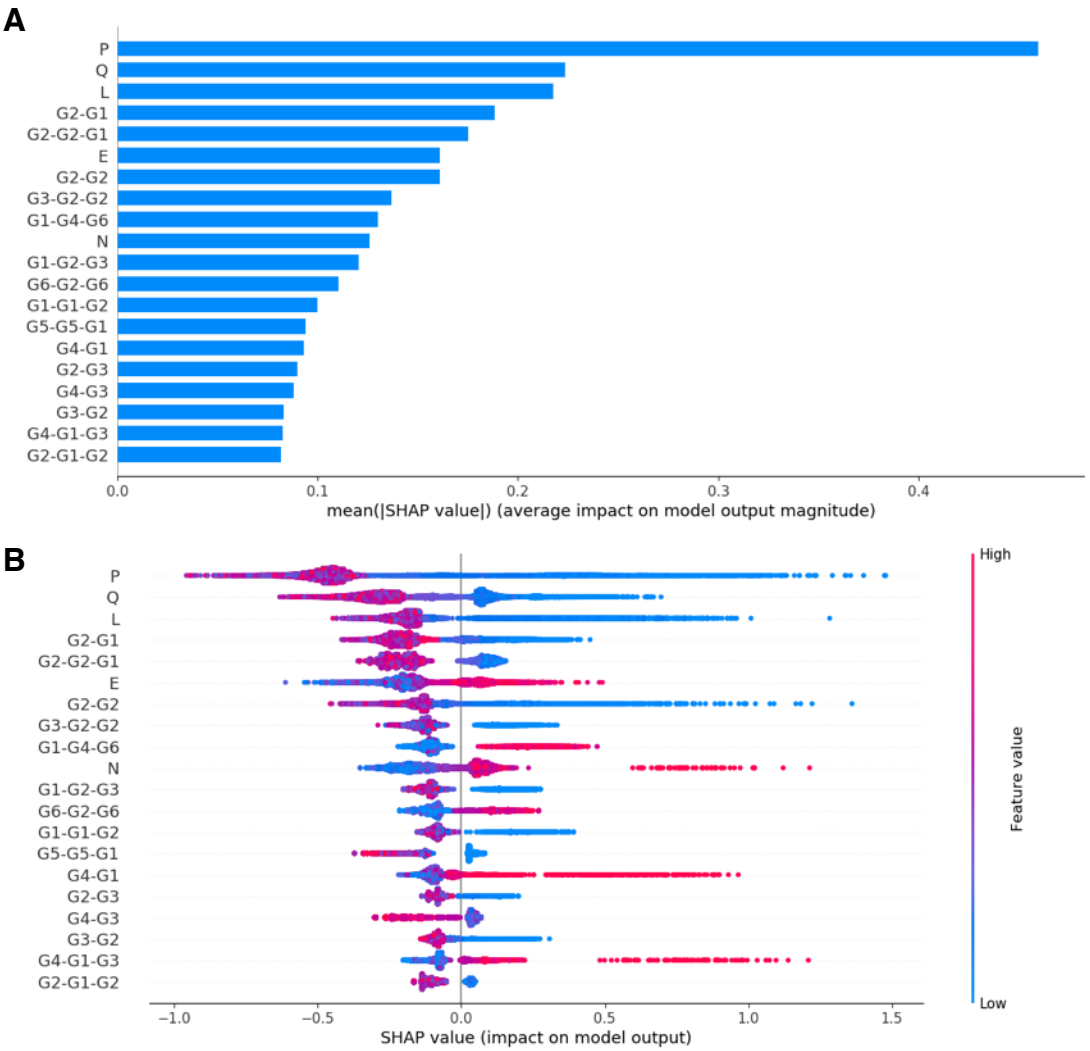


Figure S1. SHAP analysis of AcRanker features. (**A**) Absolute mean of the SHAP (SHapley Additive exPlanations) (1) values as measured for the 20 highest impact features in the AcRanker model. G1-G6 represent amino acid groupings used for computing dimeric and trimeric frequencies in AcRanker. Individual amino acids are grouped according to their side-chain volume and dipole moment (Table S3) (2). (**B**) Violin plots showing the SHAP value vs. the feature value for the 20 highest impact features in AcRanker. Higher feature values (red) with negative SHAP values indicate features that tend to be absent in the training set anti-CRISPRs, while high measured feature values with positive SHAP values suggest features that are more frequently found in the training set anti-CRISPRs. The data suggest that candidates with lower proline (P), glutamine (Q), and leucine (L) content will tend to have higher rankings.

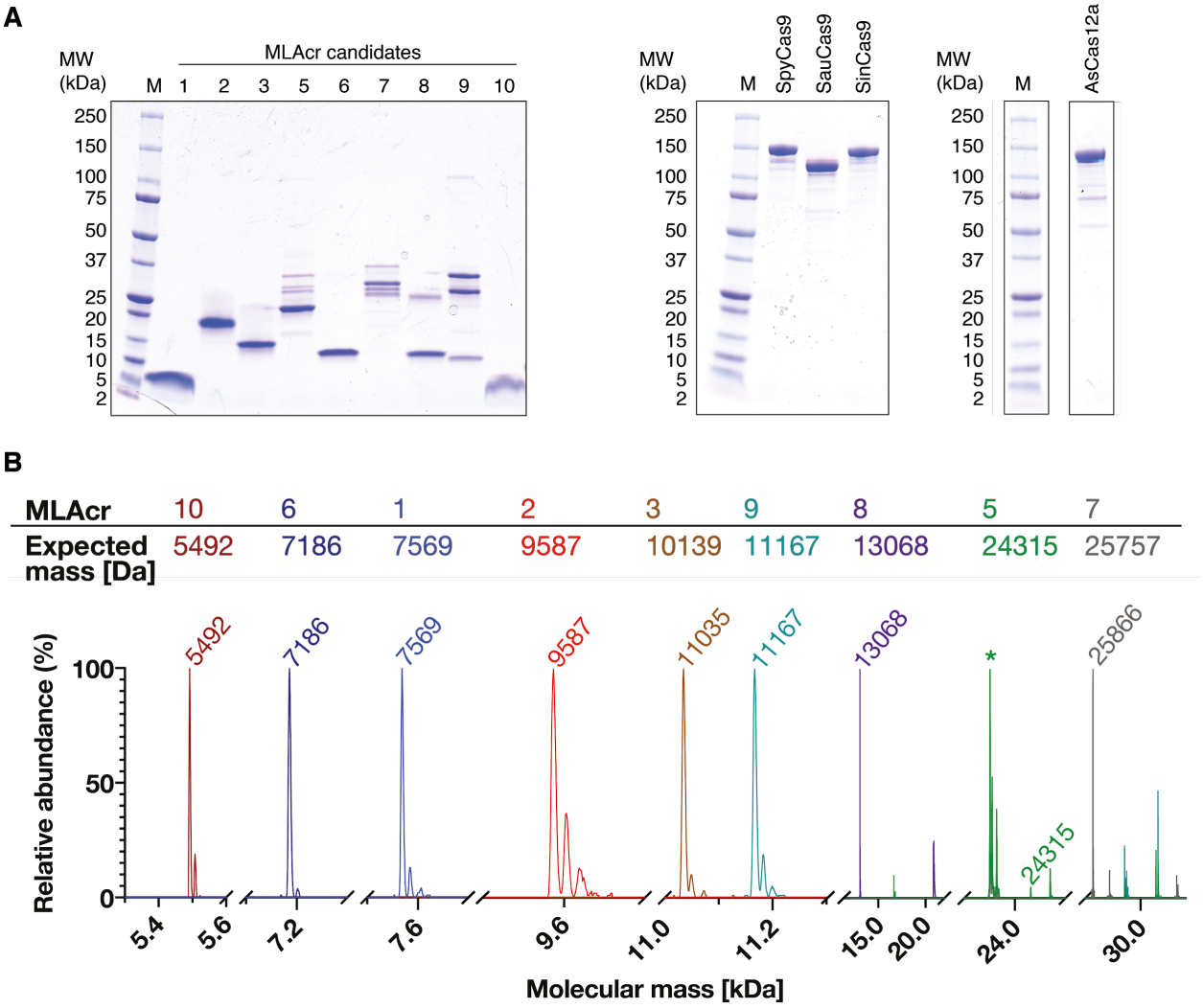


Figure S2. Purified Acr candidates and Cas effectors used in this study. (**A**) 4-20% gradient SDS-PAGE showing a size marker (M) and (left to right) purified machine learning Acr candidates, Cas9 effectors and AsCas12a used in this study. (**B**) Mass spectra of each purified Acr candidate used in this study. The measured mass of ML3 is 896 Da higher than the expected mass. We did not investigate the mass difference any further. ML5 contained a significant unidentified contaminant (*) of 23,510 Da in size.

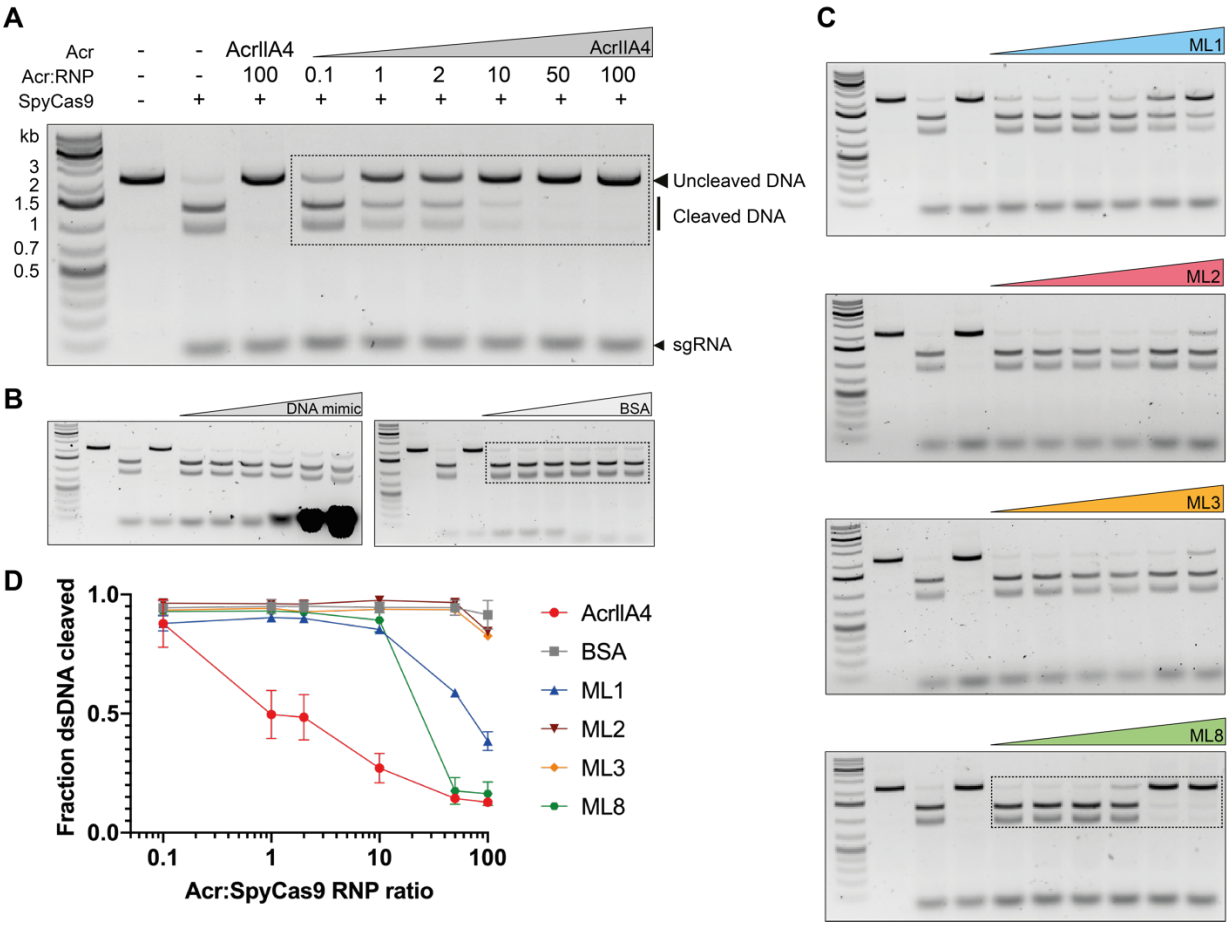


Figure S3. **Inhibition of SpyCas9 by newly discovered Acr candidates.** (**A**) *In vitro* cleavage of dsDNA by SpyCas9 in the presence of increasing concentrations of AcrllA4 (positive control). (**B**) *In vitro* cleavage of dsDNA by SpyCas9 in the presence of increasing concentrations of (left) DNA mimic and (right) BSA (DNA or BSA:RNP 0.1-, 1-, 2- ,10-, 50- and 100-fold excess from left to right). (**C**) *In vitro* cleavage of dsDNA by SpyCas9 in the presence of increasing concentrations of ML1, ML2, ML3 and ML8 (Acr:RNP 0.1-, 1-, 2- ,10-, 50- and 100-fold excess from left to right). (**D**) Quantified band intensities of the *in vitro* cleavage assays. Fraction of dsDNA cleaved (y-axis) is plotted against the Acr to SpyCas9 RNP ratio (x-axis). AcrllA4, BSA, ML1 and ML8 were run in triplicates.

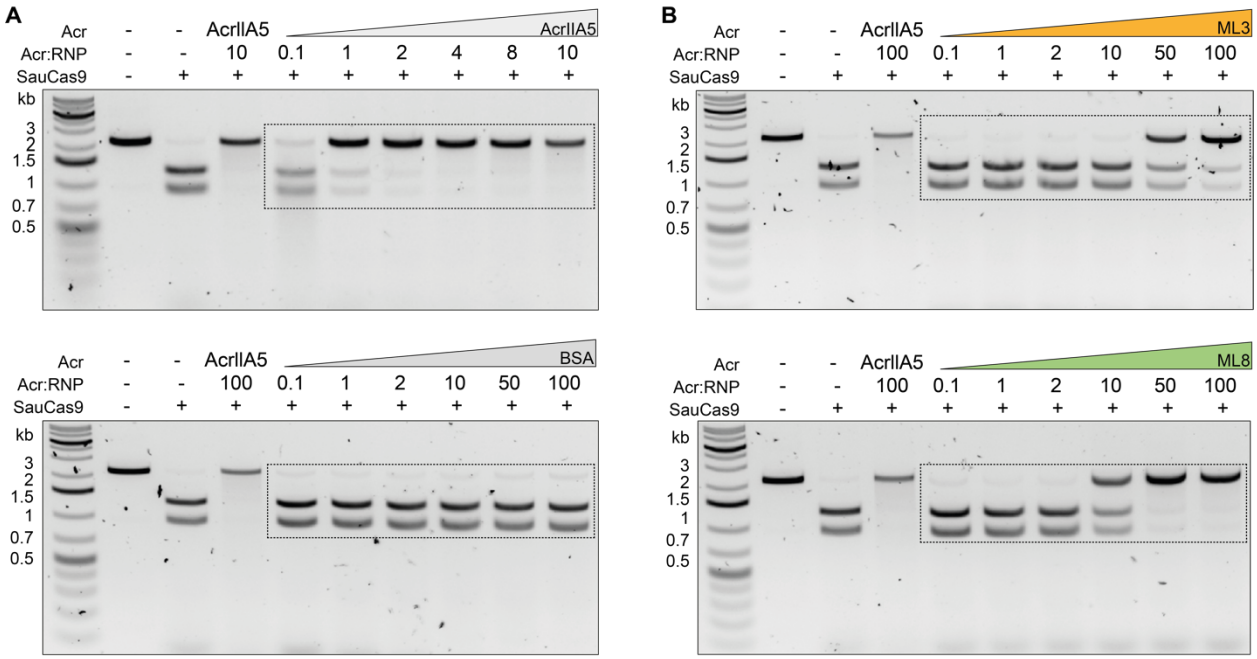


Figure S4. Inhibition of SauCas9 by newly discovered Acr candidates. (**A**) *In vitro* cleavage of dsDNA by SauCas9 in the presence of increasing concentrations of the positive control AcrllA5 (top) or negative control BSA (bottom). (**B**) *In vitro* cleavage of dsDNA by SauCas9 in the presence of increasing concentrations of ML3 (top) and ML8 (bottom).

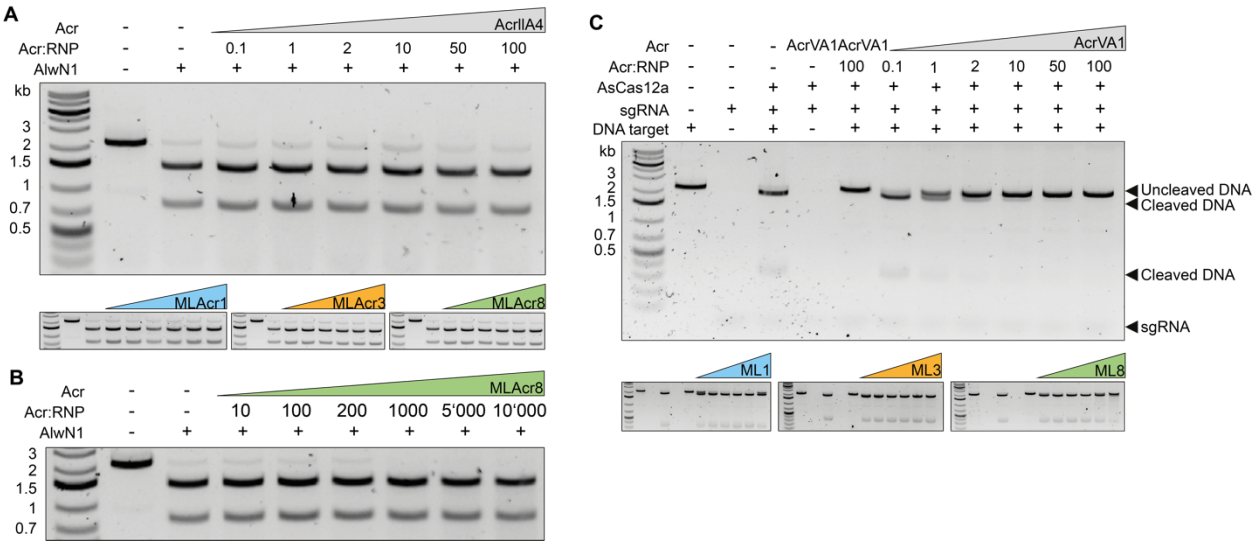


Figure S5. Control experiments for *in vitro* dsDNA cleavage assay. (**A**) *In vitro* cleavage of dsDNA by the restriction enzyme AlwN1 in the absence or presence of increasing concentrations of AcrllA4, ML1, ML3 and ML8 (Acr:AlwN1 0.1-, 1-, 2- ,10-, 50- and 100-fold excess from left to right). (**B**) *In vitro* cleavage of dsDNA by the restriction enzyme AlwN1 in the presence of increasing concentrations of ML8. (**C**) *In vitro* cleavage of dsDNA by AsCas12a in the absence or presence of increasing concentrations of AcrVA1, ML1, ML3 and ML8 (Acr:RNP 0.1-, 1-, 2- ,10-, 50- and 100-fold excess from left to right).

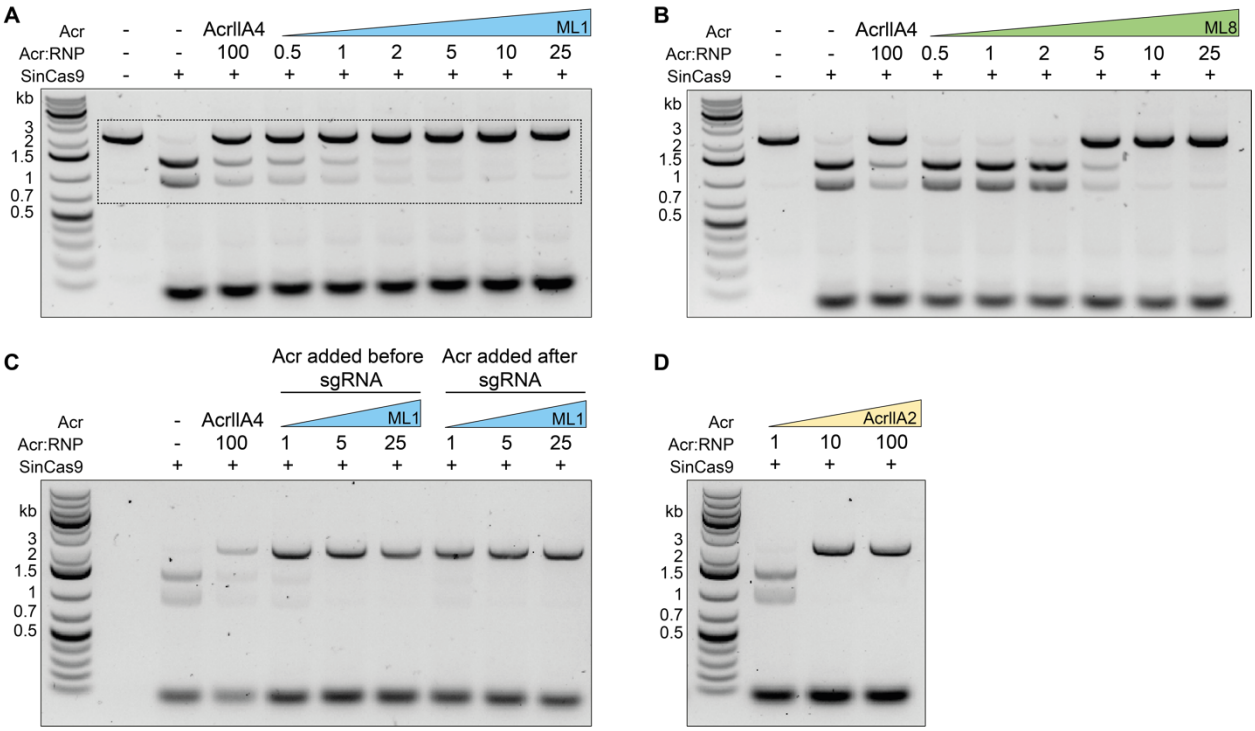


Figure S6. Inhibition of SinCas9 by ML1, ML8 and AcrIIA2. (**A**) *In vitro* cleavage of dsDNA by SinCas9 in the absence or presence of increasing concentrations of ML1. (**B**) *In vitro* cleavage of dsDNA by SinCas9 in the absence or presence of increasing concentrations of ML8. (**C**) *In vitro* cleavage assay where ML1 is incubated with SinCas9 before and after the incubation with sgRNA. (**D**) *In vitro* cleavage of dsDNA by SinCas9 in the presence of increasing concentrations of AcrIIA2. The same DNA target is used in all gels.

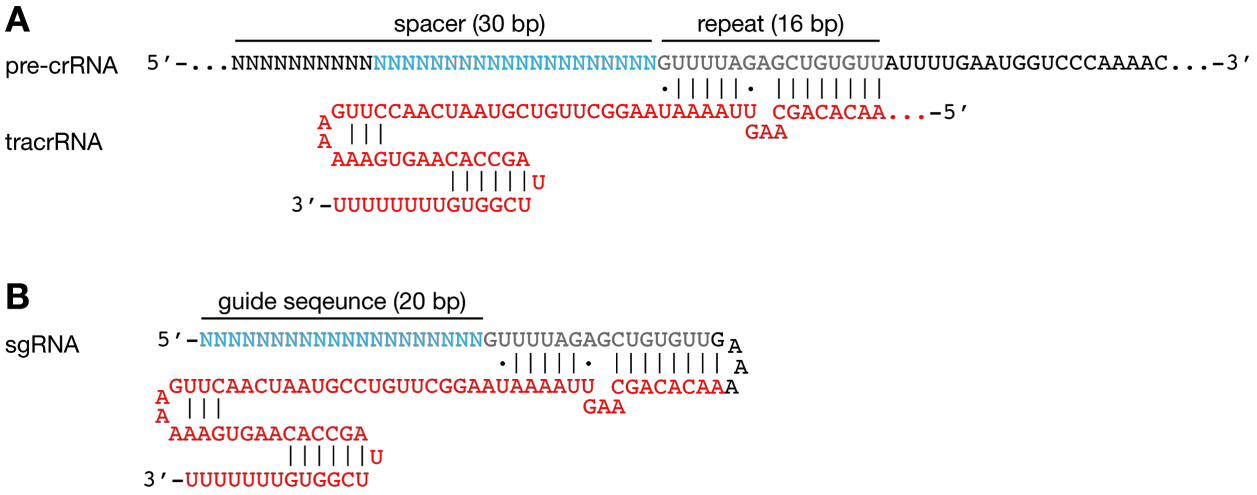


Figure S7. Design of the SinCas9 single guide RNA. (**A**) Predicted structure of the SinCas9 pre-crRNA:tracrRNA complex. (**B**) Schematic of the single-guide RNA designed for use in this study. The 20-bp spacer sequence is shown in blue, tracrRNA is shown in red and the direct repeat sequence in gray.

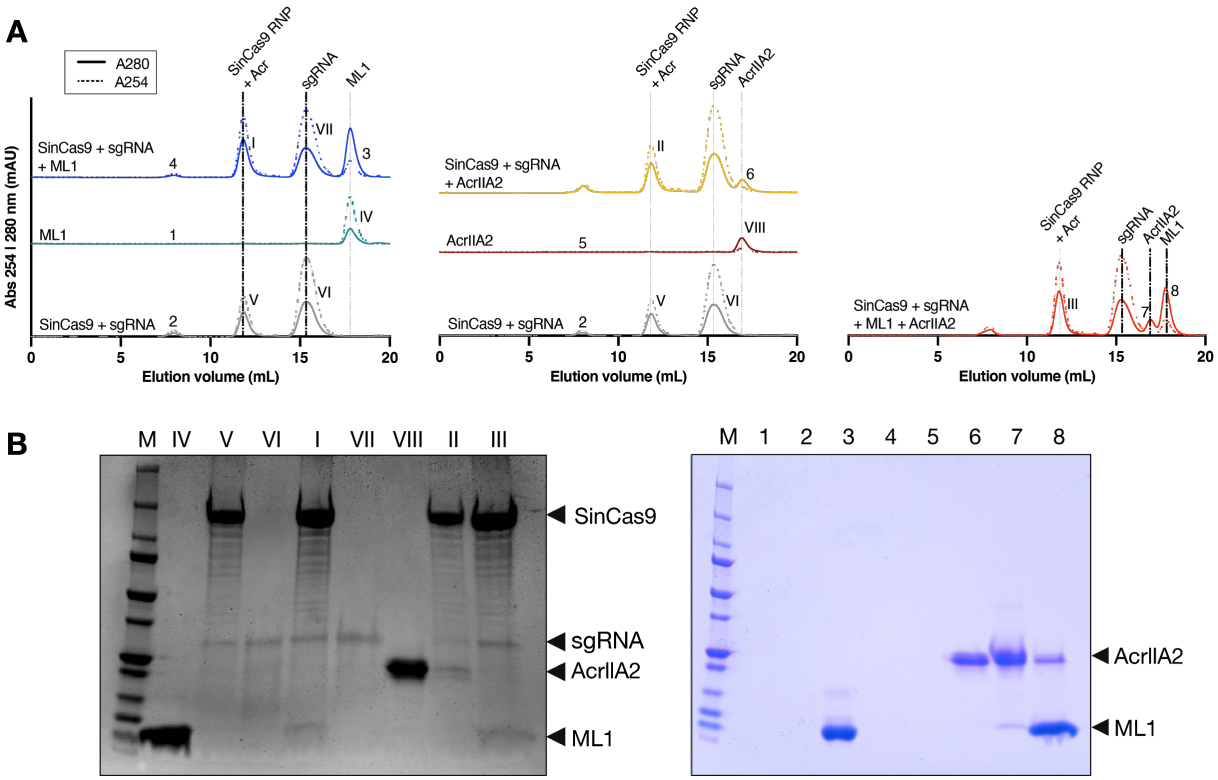


Figure S8. Competition binding experiment between ML1 and AcrIIA2. (**A**) Size-exclusion chromatogram of SinCas9-sgRNA in the absence or presence of ML1 (left), AcrIIA2 (middle) or both (right). (**B**) Coomassie-stained polyacrylamide gel illustrating the components of the fractions annotated with (I) to (VI) and 1 to 8 in panel (A).

Table S1. List of Acrs used for training and cross-validation of the AcRanker model.

Anti-CRISPRdb Name	Acr Family	Protein Accession #	Species	Proteome Size	Ref
anti_CRISPR0407	AcrIE1	YP_007392738.1	*Pseudomonas* phage JBD5	57	(3)
anti_CRISPR0408	AcrIE3	YP_950454.1	*Pseudomonas* phage DMS3	52	(3)
anti_CRISPR0409	AcrIE2	YP_007392439.1	*Pseudomonas* phage JBD88a	54	(3)
anti_CRISPR0410	AcrIE4	NP_938238.1	*Pseudomonas* phage D3112	54	(3)
anti_CRISPR0001	AcrIF1	YP_007392342.1	*Pseudomonas* phage JBD30	56	(4)
anti_CRISPR0007	AcrIF2	YP_002332454.1	*Pseudomonas* phage MP29	51	(4)
anti_CRISPR0003	AcrIF3	YP_007392440.1	*Pseudomonas* phage JBD88a	54	(4)
anti_CRISPR0002	AcrIF4	YP_007392799.1	*Pseudomonas* phage JBD24	57	(4)
anti_CRISPR0005	AcrIF5	YP_007392740.1	*Pseudomonas* phage JBD5	57	(4)
anti_CRISPR0008	AcrIF6	WP_043884810.1	*Pseudomonas aeruginosa*	6095	(5)
anti_CRISPR0011	AcrIF6	WP_019933870.1	*Oceanimonas smirnovii*	3045	(5)
anti_CRISPR0013	AcrIF6	WP_014702809.1	*Methylophaga frappieri*	2689	(5)
anti_CRISPR0022	AcrIF7	ACD38920.1	*Pseudomonas aeruginosa strain PACS458 clone fa1376 Pseudomonas aeruginosa*	57	(5)
anti_CRISPR0034	AcrIF8	AFC22483.1	*Pectobacterium* phage ZF40	68	(5)
anti_CRISPR0038	AcrIF9	WP_031500045.1	*Vibrio parahaemolyticus*	4928	(5)
anti_CRISPR0051	AcrIF10	KEK29119.1	*Shewanella xiamenensis*	3552	(5)
anti_CRISPR0134	AcrIIA1	AEO04364.1	*Listeria monocytogenes* J0161	2952	(6)
anti_CRISPR0246	AcrIIA2	AEO04363.1	*Listeria monocytogenes* J0161	2952	(6)
anti_CRISPR0384	AcrIIA4	AEO04689.1	*Listeria monocytogenes* J0161	2952	(6)
anti_CRISPR0433	AcrIIA5	D4276_028	*Streptococcus* phage D4276	54	(7)

Table S2. List of Acrs used for independent testing of AcRanker.

Acr Family	Protein Accession	Species	Proteome Size	Ref
AcrIE5	WP_074973300.1	*Pseudomonas otitidis* strain DSM 17224	5731	(8)
AcrIE6	WP_087937214.1	*Pseudomonas aeruginosa* strain S708_C14_RS	6794	(8)
AcrIE7	WP_087937215.1	*Pseudomonas aeruginosa* strain S708_C14_RS	6794	(8)
AcrIE4-F7	WP_064584002.1	*Pseudomonas citronellolis* strain SJTE-3	6260	(8)
AcrIF11	WP_038819808.1	*Pseudomonas aeruginosa* str. C1426	5888	(8)
AcrIF11.1	WP_033936089.1	*Pseudomonas aeruginosa* strain TRN6649	6373	(8)
AcrIF11.2	EGE18857.1	*Moraxella catarrhalis* BC8	1844	(8)
AcrIF12	ABR13388.1	*Pseudomonas aeruginosa* PAGI-5 genomic island sequence	121	(8)
AcrIF13	EGE18854.1	*Moraxella catarrhalis* BC8	1843	(8)
AcrIF14	AKI27193.1	*Moraxella* phage Mcat5	68	(8)
AcrIC1	WP_046701304.1	*Moraxella bovoculi strain 58069*	1944	(8)
AcrIIA3	WP_014930691.1	*Listeria monocytogenes serotype 7 str. SLCC2482*	2822	(6)
AcrIIA6	WP_149028791.1	*Streptococcus phage D1811*	40	(9)
AcrIIA7	AII65827.1	*Bacteroides dorei isolate HS1_L_1_B_010*	4519	(10)
AcrIIA9	WP_004289410.1	*Bacteroides fragilis strain DCMOUH0067B*	4286	(10)
AcrIIA13	AKS70260.1	*Staphylococcus schleiferi strain 5909-02*	2278	(11)
AcrIIC5	WP_002642161.1	*Simonsiella muelleri ATCC 29453*	2170	(12)
AcrIIIB1	NP_666582.1	*Sulfolobus islandicus rod-shaped virus 2*	54	(13)
AcrVA1	WP_046701302.1	*Moraxella bovoculi strain 58069*	1944	(8, 14)
AcrVA4	WP_046699156.1	*Moraxella bovoculi strain 22581*	2105	(14)

Table S3. Grouping of amino acids based on physiochemical properties. Groups of amino acids with similar side chains are grouped together to reduce the number of features to test in the machine learning model (2).

Group #	Dipole Scale[a]	Volume Scale[b]	Amino Acids
1	-	-	A, G, V
2	-	+	I, L, F, P
3	+	+	Y, M, T, S
4	++	+	H, N, Q, W
5	+++	+	R, K
6	+'+'+'	+	D, E
7	+[c]	+	C

[a] Dipole scale (Debye): -, Dipole < 1.0; +, 1.0 < Dipole < 2.0; ++, 2.0 < Dipole < 3.0; +++, Dipole > 3.0; +'+'+', Dipole > 3.0 with opposite orientation

[b] Volume scale ($Å^3$): -, Volume < 50; +, Volume > 50

[c] Cysteine is separated from class 3 because of its ability to form disulfide bonds

Table S4. Comparison of XGBoost classification vs. pairwise ranking models during leave-one-out cross-validation. Each row of the table indicates which Acr was excluded from the training dataset and used as a test dataset, with the number indicating the rank obtained using either a blastp search against all other known Acrs in the training set (blastp), an XGBoost classification model (Class.), an XGBoost pairwise ranking model (Ranking). The best rank achieved by the XGBoost classification or pairwise ranking model within the complete or prophage proteome is marked with an asterisk. The best rank between blastp and either XGBoost model is bolded, and any method that produces the top rank is bolded with two asterisks. The pairwise ranking model performs better than the classification model, with the ranking model receiving a better rank 11 times vs. six times for the classification model in complete bacterial or phage proteomes. In the smaller prophage proteomes the ranking model is ranked higher five times vs. once for the classification model.

Protein	Acr Family	Complete Proteome				Prophage Proteome Subset			
		Size	blastp	AcRanker (XGBoost)		Size	blastp	AcRanker (XGBoost)	
				Class.	Ranking			Class.	Ranking
anti_CRISPR0407	AcrIE1	57	33	9	**1****				
anti_CRISPR0408	AcrIE3	52	17	**1****	**1****				
anti_CRISPR0409	AcrIE2	54	18	5	**2***				
anti_CRISPR0410	AcrIE4	54	11	2	**1****				
anti_CRISPR0001	AcrIF1	56	21	**4***	11				
anti_CRISPR0007	AcrIF2	51	34	**1****	**1****				
anti_CRISPR0003	AcrIF3	54	5	9	**1****				
anti_CRISPR0002	AcrIF4	57	36	**1****	3				
anti_CRISPR0005	AcrIF5	57	26	**19***	**19***				
anti_CRISPR0008	AcrIF6	6095	**1****	69*	80	361	**1****	17	15*
anti_CRISPR0011	AcrIF6	3045	**1****	25	13*	72	**1****	3	**1****
anti_CRISPR0013	AcrIF6	2689	**1****	541	130*	57	-	-	-
anti_CRISPR0022	AcrIF7	57	20	3	**1****				
anti_CRISPR0034	AcrIF8	68	30	3	**1****				
anti_CRISPR0038	AcrIF9	4928	198	**44***	333	37	-	-	-
anti_CRISPR0051	AcrIF10	3552	189	**2***	17	70	23	**1****	2
anti_CRISPR0134	AcrIIA1	2951	183	931	**770***	146	**60**	97	87*
anti_CRISPR0246	AcrIIA2	2952	210	15*	16	146	34	6	**3***
anti_CRISPR0384	AcrIIA4	2951	59	56	**21***	146	9	15	**4***
anti_CRISPR0433	AcrIIA5	54	**5**	12	**8***				

Table S5. Amino acid sequence and accession numbers of all the Acr candidates.

#ML cand.	Accession No.	Sequence
ML1	OHX26873.1	MKNYEVTNEVKNLNTQVETIGQAVDLYKEYGSNTIVWSIDK NEDLIDEVTELVAEYAEKGTVIK
ML2	WP_003731277.1	MGKTYWYNEGTDTLLTEKEYKELMEREAKALYEEVQEEEKD FESSEKTSFEEFLKTCYENESDFVLSDNEGNKLEEW
ML3	WP_003731276.1	MSKTMYKNDVIELIKNAKTNNEELLFTSVERNTREAATQYFR CPEKHVSDAGVYYGEDFEFDGFEIFEDDLIYTRSYDKEELN
ML4	WP_000946250.1	MLRRVNHVKNVLAHGEFAEWIENKIGIHYREANRMMTVAKQ IPNVSTLKYLGATAKHVNGVAKRKQNFLSQISLIPTNPQLPHQ TIINTYLYWQP
ML5	WP_001080841.1	MNRLKELRKEKKLTQEELAGEIGVSKITILRWENGERQIKPDK AKELAKYFNVSVGYLLGYAPNKKIDFQLNLDGTTLHLTKEQF LALENTSKSIKKIKNTINESVKQEEYIKNASKYYDFEKVSRRLT DRLFEIHTDLIELLMMLDHFPSGELSKSQQEAIFKFYKQLDYFV TDTPASFDYFKKNLESYGYKIYTEGDKIDFD
ML6	WP_000965633.1	MLYIDEFKEAIDKGYILGGTVAIVRKNGKIFDYVLPHEEVREE EVVTVERVEDVMRELE
ML7	WP_000591144.1	MIKIYFGKDAALNQAIQSRLDSYQIDYQAFSSKDIDAKTLMEW LFKSTDIFELLSTKMLKYKLNTQITLSQFVRKILKDVNSTLKLPI VVTDEVIYSNMSPDYVTVLLPKEYRKIKRIQLMRKMEQLDEG RLFWKNFELFRKQSELRWFELNELLFADVSDDLGEIKKAKDR FFSYKKNNQVPPNEIIERILKIFLVDREDFFKKSPSDLQNF
ML8	WP_000384271.1	MDYDNENYLIPKILLQDDFYSSLSAKDILVYAVLKDRQIEALE KGWIDTDGSIYLNFKLIELAKMFSCSRTTMIDVMQRLEEVNLI ERERVDVFYGYSLPYKTYINEV
ML9	WP_000134666.1	MTEGFTIQLPKVTEKKLLARYDDMLQKAIEKALEDKELYKPI VRMAGLCRWLDVSTTTVVKWQKQGGMPHMVIDGVTLYDK HKVAQWLQQFER
ML10	WP_011058321.1	MNIEDIERIISEYLIFRSDIDGCAVIDIEDFLKHIRFSYERLK

Table S6. Amino acid sequence of all the Cas effectors used in this study.

Cas effector	Species	Sequence
Cas9	*Streptococcus pyogenes*	MDKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIK KNLIGALLFDSGETAEATRLKRTARRRYTRRKNRICYLQEIFSNE MAKVDDSFFHRLEESFLVEEDKKHERHPIFGNIVDEVAYHEKYP TIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDN SDVDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLE NLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQLSK DTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSDILRVNTEIT KAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKN GYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRK QRTFDNGSIPHQIHLGELHAILRRQEDFYPFLKDNREKIEKILTFRI PYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKGASAQSFI ERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMR KPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEI SGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLF EDREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGI RDKQSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDIQKAQVS GQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPEN IVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVEN TQLQNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDHIVPQSF LKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMKNYWRQLLN AKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVA QILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVR EINNYHHAHDAYLNAVVGTALIKKYPKLESEFVYGDYKVYDVR KMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLANGEIRKRPLIET NGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKE SILPKRNSDKLIARKKDWDPKKYGGFDSPTVAYSVLVVAKVEKG KSKKLKSVKELLGITIMERSSFEKNPIDFLEAKGYKEVKKDLIIKL PKYSLFELENGRKRMLASAGELQKGNELALPSKYVNFLYLASHY EKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANL DKVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTID RKRYTSTKEVLDATLIHQSITGLYETRIDLSQLGGD
Cas9	*Staphylococcus aureus*	MGKRNYILGLDIGITSVGYGIIDYETRDVIDAGVRLFKEANVENN EGRRSKRGARRLKRRRRHRIQRVKKLLFDYNLLTDHSELSGINP

		YEARVKGLSQKLSEEEFSAALLHLAKRRGVHNVNEVEEDTGNE LSTKEQISRNSKALEEKYVAELQLERLKKDGEVRGSINRFKTSDY VKEAKQLLKVQKAYHQLDQSFIDTYIDLLETRRTYYEGPGEGSP FGWKDIKEWYEMLMGHCTYFPEELRSVKYAYNADLYNALNDL NNLVITRDENEKLEYYEKFQIIENVFKQKKKPTLKQIAKEILVNEE DIKGYRVTSTGKPEFTNLKVYHDIKDITARKEIIENAELLDQIAKI LTIYQSSEDIQEELTNLNSELTQEEIEQISNLKGYTGTHNLSLKAIN LILDELWHTNDNQIAIFNRLKLVPKKVDLSQQKEIPTTLVDDFILS PVVKRSFIQSIKVINAIIKKYGLPNDIIIELAREKNSKDAQKMINEM QKRNRQTNERIEEIIRTTGKENAKYLIEKIKLHDMQEGKCLYSLE AIPLEDLLNNPFNYEVDHIIPRSVSFDNSFNNKVLVKQEENSKKG NRTPFQYLSSSDSKISYETFKKHILNLAKGKGRISKTKKEYLLEER DINRFSVQKDFINRNLVDTRYATRGLMNLLRSYFRVNNLDVKVK SINGGFTSFLRRKWKFKKERNKGYKHHAEDALIIANADFIFKEW KKLDKAKKVMENQMFEEKQAESMPEIETEQEYKEIFITPHQIKHI KDFKDYKYSHRVDKKPNRELINDTLYSTRKDDKGNTLIVNNLN GLYDKDNDKLKKLINKSPEKLLMYHHDPQTYQKLKLIMEQYGD EKNPLYKYYEETGNYLTKYSKKDNGPVIKKIKYYGNKLNAHLDI TDDYPNSRNKVVKLSLKPYRFDVYLDNGVYKFVTVKNLDVIKK ENYYEVNSKCYEEAKKLKKISNQAEFIASFYNNDLIKINGELYRV IGVNNDLLNRIEVNMIDITYREYLENMNDKRPPRIIKTIASKTQSI KKYSTDILGNLYEVKSKKHPQIIKKG
Cas9	*Streptococcus iniae*	MRKPYSIGLDIGTNSVGWAVITDDYKVPSKKMRIQGTTDRTSIK KNLIGALLFDNGETAEATRLKRTTRRRYTRRKYRIKELQKIFSSE MNELDIAFFPRLSESFLVSDDKEFENHPIFGNLKDEITYHNDYPTI YHLRQTLADRDQKADLRLIYLALAHIIKFRGHFLIEGNLDSENTD VHVLFLNLVNIYNNLFEEDIVETASIDAEKILTSKTSKSRRLENLIA EIPNQKRNMLFGNLVSLALGLTPNFKTNFELLEDAKLQISKDSYE EDLDNLLAQIGDQYADLFIAAKKLSDAILLSDIITVKGASTKAPLS ASMVQRYEEHQQDLALLKNLVKKQIPEKYKEIFDNKEKNGYAG YIDGKTSQEEFYKYIKPILLKLNGTEKLISKLEREDFLRKQRTFDN GSIPHQIHLNELKAIIRRQEKFYPFLKENQKKIEKLFTFKIPYYVGP LANGQSSFAWLKRQSNESITPWNFEEVVDQEASARAFIERMTNF DTYLPEEKVLPKHSPLYEMFMVYNELTKVKYQTEGMKRPVFLS SEDKEEIVNLLFKKDRKVTVKQLKEEYFSKMKCFHTVTILGVED RFNASLGTYHDLLKIFKDKAFLDDEANQDILEEIVWTLTLFEDQA

		MIERRLVKYADVFEKSVLKKLKKRHYTGWGRLSQKLINGIKDK QTGKTILGFLKDDGVANRNFMQLINDSSLDFAKIIKHEQEKTIKN ESLEETIANLAGSPAIKKGILQSIKIVDEIVKIMGQNPDNIVIEMAR ENQSTMQGIKNSRQRLRKLEEVHKNTGSKILKEYNVSNTQLQSD RLYLYLLQDGKDMYTGKELDYDNLSQYDIDHIIPQSFIKDNSIDN IVLTTQASNRGKSDNVPNIEIVNKMKSFWYKQLKNGAISQRKFD HLTKAERGALSDFDKAGFIKRQLVETRQITKHVAQILDSRFNSNL TEDSKSNRNVKIITLKSKMVSDFRKDFGFYKLREVNDYHHAQDA YLNAVVGTALLKKYPKLEAEFVYGDYKHYDLAKLMIQPDSSLG KATTRMFFYSNLMNFFKKEIKLADDTIFTRPQIEVNTETGEIVWD KVKDMQTIRKVMSYPQVNIVMKTEVQTGGFSKESILPKGNSDKL IARKKSWDPKKYGGFDSPIIAYSVLVVAKIAKGKTQKLKTIKELV GIKIMEQDEFEKDPIAFLEKKGYQDIQTSSIIKLPKYSLFELENGRK RLLASAKELQKGNELALPNKYVKFLYLASHYTKFTGKEEDREK KRSYVESHLYYFDEIMQIIVEYSNRYILADSNLIKIQNLYKEKDNF SIEEQAINMLNLFTFTDLGAPAAFKFFNGDIDRKRYSSTNEIINSTL IYQSPTGLYETRIDLSKLGGK
Cas12a	*Acidaminococcus sp.*	MTQFEGFTNLYQVSKTLRFELIPQGKTLKHIQEQGFIEEDKARND HYKELKPIIDRIYKTYADQCLQLVQLDWENLSAAIDSYRKEKTEE TRNALIEEQATYRNAIHDYFIGRTDNLTDAINKRHAEIYKGLFKA ELFNGKVLKQLGTVTTTEHENALLRSFDKFTTYFSGFYENRKNV FSAEDISTAIPHRIVQDNFPKFKENCHIFTRLITAVPSLREHFENVK KAIGIFVSTSIEEVFSFPFYNQLLTQTQIDLYNQLLGGISREAGTEK IKGLNEVLNLAIQKNDETAHIIASLPHRFIPLFKQILSDRNTLSFILE EFKSDEEVIQSFCKYKTLLRNENVLETAEALFNELNSIDLTHIFISH KKLETISSALCDHWDTLRNALYERRISELTGKITKSAKEKVQRSL KHEDINLQEIISAAGKELSEAFKQKTSEILSHAHAALDQPLPTTLK KQEEKEILKSQLDSLLGLYHLLDWFAVDESNEVDPEFSARLTGIK LEMEPSLSFYNKARNYATKKPYSVEKFKLNFQMPTLASGWDVN KEKNNGAILFVKNGLYYLGIMPKQKGRYKALSFEPTEKTSEGFD KMYYDYFPDAAKMIPKCSTQLKAVTAHFQTHTTPILLSNNFIEPL EITKEIYDLNNPEKEPKKFQTAYAKKTGDQKGYREALCKWIDFT RDFLSKYTKTTSIDLSSLRPSSQYKDLGEYYAELNPLLYHISFQRI AEKEIMDAVETGKLYLFQIYNKDFAKGHHGKPNLHTLYWTGLF SPENLAKTSIKLNGQAELFYRPKSRMKRMAHRLGEKMLNKKLK DQKTPIPDTLYQELYDYVNHRLSHDLSDEARALLPNVITKEVSHE

		IIKDRRFTSDKFFFHVPITLNYQAANSPSKFNQRVNAYLKEHPETP IIGIDRGERNLIYITVIDSTGKILEQRSLNTIQQFDYQKKLDNREKE RVAARQAWSVVGTIKDLKQGYLSQVIHEIVDLMIHYQAVVVLE NLNFGFKSKRTGIAEKAVYQQFEKMLIDKLNCLVLKDYPAEKVG GVLNPYQLTDQFTSFAKMGTQSGFLFYVPAPYTSKIDPLTGFVDP FVWKTIKNHESRKHFLEGFDFLHYDVKTGDFILHFKMNRNLSFQ RGLPGFMPAWDIVFEKNETQFDAKGTPFIAGKRIVPVIENHRFTG RYRDLYPANELIALLEEKGIVFRDGSNILPKLLENDDSHAIDTMV ALIRSVLQMRNSNAATGEDYINSPVRDLNGVCFDSRFQNPEWPM DADANGAYHIALKGQLLLNHLKESKDLKLQNGISNQDWLAYIQ ELRN

Table S7. sgRNAs used for the *in vitro* cleavage assay.

Cas effector	**Species**	**sgRNA sequence***	**Main Text Fig.**	**Suppl. Fig.**
Cas9	*Streptococcus pyogenes*	**ATACGGGAGGGCTTACCATC**GTTTTAGAGCTATGCTGTTTTGGAAACAAAACAGCATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT	2A-B	3A-D
Cas9	*Staphylococcus aureus*	**TATCGTAGTTATCTACACGACG**GTTTTAGTACTCTGGAAACAGAATCTACTAAAACAAGGCAAAATGCCGTGTTTATCTCGTCAACTTGTTGGCGAGATTTTT	2C-D	4A-B
Cas9	*Streptococcus iniae*	**ATACGGGAGGGCTTACCATC**GTTTTAGAGCTGTGTTGAAAAACACAGCAAGTTAAAATAAGGCTTGTCCGTAATCAACTTGAAAAAGTGAACACCGATTCGGTGTTTTTTT	3A-B	6A-D
Cas12a	*Acidaminococcus sp.*	AAUUUCUACUCUUGUAGAU**AAAGUGCUCAUCAUUGGAAAACGU**	-	5C

* Spacer sequences are shown in bold

Table S8. DNA target used for the *in vitro* cleavage assay.

Cas effector	Species	DNA target sequence*	Main Text Fig.	Suppl. Fig.
Cas9	*Streptococcus pyogenes*	AATAATGGTTTCTTAGACGTCAGGTGGCACTTTTCGGGGA AATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATAC ATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATA AATGCTTCAATAATATTGAAAAAGGAAGAGTATGAGTAT TCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCAT TTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAA AGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGG TTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGA GAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACT TTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTG ACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATT CTCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAA AGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCA GTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACT TACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCG CTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGA TCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGA CGAGCGTGACACCACGATGCCTGTAGCAATGGCAACAAC GTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCT TCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAA GTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCT GGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTC TCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCC CTCCCG**TAT**CGTAGTTATCTACACGACGGGGAGTCAGG CAACTATGGATGAACGAAATAGACAGATCGCTGAGATAG GTGCCTCACTGATTAAGCATTGGTAACTGTCAGACCAAGT TTACTCATATATACTTTAGATTGATTTAAAACTTCATTTTT AATTTAAAAGGATCTAGGTGAAGATCCTTTTTGATAATCT CATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGA GCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGA GATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAA AAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATC AAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAG CAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCC GTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCT ACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTG	2A-B	3A-D
Cas9	***Staphylococcus aureus***		2C-D	4A-B
Cas9	*Streptococcus iniae* UEL-Si1		3A-B	6A-D
Cas12a	*Acidaminococcus sp.*		-	5C

		CCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAA GACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAA CGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGA CCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAG AAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGG TATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCAC GAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAG TCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTT TTGTGATGCTCGTCAGGGGGGCGGAGCCTGTGGAAAAAC GCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCT GGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGAT TCTGTGGATAACCGTATTACCGCAGAGTTTGTAGAAACGC AAAAAGGCCATCCGTCAGGATGGCCTTCTGCTTAATTTGA TGCCTGGCAGTTTATGGCGGGCGTCCTGCCCGCCACCCTC CGGGCCGTTGCTTCGCAACGTTCAAATCCGCTCCCGGCGG TTGAGAAGAGAAAAGAAAACCGCCGATCCTGTCCACCGC ATTACTGCAAGGTAGTGGACAAGACCGGCGGTCTTAAGT TTTTTGGCTGAAATGCCTGGCAGTTCCCTACTCTCGCATG GGGCTCGCGGTTAACTGATTATTTTATTTATCTAGGCTAC TTACGAACG		
		DNA mimic**	-	3B
		GCTGACAATGATACGAACGAGACACACGCTCACGACTCA G		

* Target sequences are shown in blue (*Streptococcus pyogenes, Streptococcus iniae*), green (*Acidaminococcus sp.*) or bold (*Staphylococcus aureus*); **DNA mimic used for control experiments

Table S9. Independent testing set validation results. 20 proteomes containing non-redundant (<40% sequence identity) Acrs from bacterial and phage sources were ranked using AcRanker and blastp. Bacterial proteomes that had Acrs within PHASTER-predicted prophages were also tested with a subset of the proteome containing only the prophage proteins. Cases where the top rank is returned are in boldface.

		Complete Proteome			Prophage Subset		
Acr Accession #	**Acr Family**	**Proteome size**	**AcRanker rank**	**Blastp rank**	**Proteome size**	**AcRanker rank**	**Blastp rank**
WP_064584002.1	AcrIE4-F7	6260	68	**1**	111	4	**1**
WP_074973300.1	AcrIE5	5731	10	63	-	-	-
WP_087937214.1	AcrIE6	6794	80	4383	-	-	-
WP_087937215.1	AcrIE7	6794	742	6546	-	-	-
WP_038819808.1	AcrIF11	5888	138	2995	64	3	38
WP_033936089.1	AcrIF11.1	6373	38	2293	92	**1**	38
EGE18857.1	AcrIF11.2	1844	412	90	59	30	**1**
ABR13388.1	AcrIF12	121	7	10	-	-	-
EGE18854.1	AcrIF13	1844	187	755	-	-	-
AKI27193.1	AcrIF14	68	14	3	68	14	3
WP_046701304.1	AcrIC1	1944	6	313	72	**1**	15
WP_014930691.1	AcrIIA3	2822	10	1184	74	2	40
WP_149028791.1	AcrIIA6	40	23	21	40	23	21
AII65827.1	AcrIIA7	4519	179	2208	-	-	-
WP_004289410.1	AcrIIA9	4286	53	930	-	-	-
AKS70260.1	AcrIIA13	2278	22	355	145	3	29
WP_002642161.1	AcrIIC5	2170	10	1954	367	6	237
NP_666582.1	AcrIIIB1	54	44	25	54	44	25
WP_046701302.1	AcrVA1	1944	114	376	72	10	18
WP_046699156.1	AcrVA4	2105	1100	1405	293	220	81

Table S10. List of expected lethal self-targeting *Streptococcus* genomes obtained with Self-Target Spacer Searcher (STSS). Searching *Streptococcus* assemblies from NCBI with STSS returned 385 cases of self-targeting derived from type II-A arrays representing 241 individual genomes. Of those genomes, 20 contained at least one spacer with the characteristic NRG 3′ PAM for SpyCas9, shown in the table below. Only *Streptococcus iniae* strain UEL-Si1 contains a previously discovered anti-CRISPR (AcrIIA3). Also shown in the table are the self-targeting spacers for *Listeria monocytogenes* strain R2-502, which was also ranked with AcRanker.

Target Accession#	Locus Accession#	Species/Strain	Self-Targeting Spacer Sequence(s)	3′ PAM Region	Anti-CRISPRs Present
NZ_MNAC01000031.1	NZ_MNAC01000010.1	*Streptococcus iniae* strain UEL-Si1	TTGATAAGTATAATTTCCTGTCTTTGTTTT	AGGAGTTTT	AcrIIA3 (WP_071127625.1)
NZ_MNAC01000046.1			TAAGGAATTTGAAGCAATACGTCTTAATTT	AGCAATGAC	
NZ_MNAC01000023.1			CAAAAAAGTTCGGTAACTTACGGTAACTTA	CGGTAACTT	
			TCTAAAAAATCAAAAGTTACCGTGTTACCG	TAGTTTTGA	
			AATATGACTTTTGGGAAATTAACTAATCAA	TGGCTGAAA	
			TTTTTGAGTGTACTGATGTTGCTTTTGAGC	TGGCCACTT	
NZ_MNAC01000021.1			ATAATCAATCACATTAATGCTGACATCAAC	TGGAGCAGA	
			GAGTTTAATTAAGTGACATAATATCTTCAT	CGGTTATAG	
NZ_JRLL01000002.1	NZ_JRLL01000058.1	*Streptococcus pyogenes* SS1447	TCGTCAGATTTGTCAGTATAGTAATCATCA	CGATATAAA	None
NZ_JRLL01000072.1			CTATATTGTTGAGCTGTGGGCTTTGCATAA	AGGTTTAAA	

NZ_JRLL01000026.1			GTAATAATAGCATTGCCTGTTCTATCCTGT	CGGTAGAAC	
NZ_CQAV01000003.1	NZ_CQAV01000001.1	*Streptococcus agalactiae* strain DE-NI-032	TATTTGATAGCGGTAACGGGTCATATACAA	AGGCATCTA	None
			TGGTGGTATTTATAATGTACGAGCAAATCG	AGGCGCTCC	
			ACCTTGCTCCGATGACACCATCGCGAACCT	TGGTCTAAT	
NZ_CP010449.1	NZ_CP010449.1	*Streptococcus pyogenes* strain NGAS322	ATCGTAAGGCAACAGATTATCGTAAGATCT	AGGTGTATA	None
NZ_ALQN01000014.1	NZ_ALQN01000018.1	Streptococcus agalactiae CCUG 37430	ATTTGCAACTTTCTCAAGTGTTGCGAGAGA	TGGAGAATT	None
NZ_ALQN01000018.1			GCAAGCACTAAATGAAGCTACTAGACTTAA	AGGTCGCAG	
			TAATGACATGTGGATTGATATCTCAGAGAA	CGGCGATTA	
			TGTCATTGTTAAAATCATTTGCATATTTTT	TGGATATAA	
			TACTTGACGAATTGAAGATGACGGAATTTA	TTGCTCCAC	
NZ_CPVL01000019.1	NZ_CPVL01000003.1	*Streptococcus agalactiae* strain DE-NI-007	AAGGCACGCGCAAGATGAATTCATTTCTAA	TGGCTACAC	None
			TGATGTTCTTTATCAAACATTCTAAATACT	TGGAAGCCC	
			GAGCCTTGCTTGAGTTTGTGGAGCTTTATA	GGGATGGAA	
			GTATAATTTAGTTAAGCTTAAATTTAACCA	AGGAGACGT	

NZ_ANCM01000101.1	NZ_ANCM01000101.1	*Streptococcus agalactiae* FSL S3-586	GAAAAAGGCGATGTAGCTTAGAAAGGAGAA	GGGATGGAA	None
NZ_ANCM01000006.1			GAAAAAGGCGATGTAGCTTAGAAAGGAGAA	CACCATGAA	
NZ_ANCM01000028.1			TACGAAAAGGTTGTGATAAAAGCCATATCA	TCGAGTTTG	
NZ_ALTM01000012.1	NZ_ALTM01000016.1	*Streptococcus agalactiae* GB00548	AACAACTTTCTTACAAAAGGTTCTAGTTTTCTT	TCGCAAAAC	
NZ_ALTM01000013.1			ACGCTCTGAGGCAGATGAGGAACAGGCGCA	TAGGCACCC	
NZ_ALUZ01000056.1	NZ_ALUZ01000054.1	*Streptococcus agalactiae* GB00984	TGAAAACAAGCGCAAAGCTGTCAGAAAACA	CGGAACTAA	None
			TACTTGACGAATTGAAGATGACGGAATTTA	TGGCTCCAC	
NZ_ALRF01000019.1	NZ_ALRF01000066.1	*Streptococcus agalactiae* BSU188	GAAACTTCGATTAGTTTGCGTACTCGCTCA	CGGCAAAAC	None
NZ_ANEM01000019.1	NZ_ANEM01000012.1	*Streptococcus agalactiae* MRI Z1-022	TTGCTGCTAGACCCAAACAGTTTATTTTTAG	GGCCAAAAA	None
NZ_ANEM01000074.1			TATTTCATCATAGAAAATCCTGCTAGTGGT	CGGTTATGG	
NZ_CQEL01000006.1	NZ_CQEL01000002.1	*Streptococcus agalactiae* strain DK-NI-014	ACACCTAGTTTCAAGTTTTTAGCAGATTTTTT	GGTTACATT	None
NZ_CQEL01000008.1			ACGCTCTGAGGCAGATGAGGAACAGGCGCA	TAGGCACCC	
NZ_MAWX01000026.1	NZ_MAWX01000055.1		ATTGACTGTTTACGATTTCCTTCCACCGTT	GGGTACAAA	None

		Streptococcus agalactiae strain DK-PW-096	TGATGAGATTTTTAAAAGACTCACTGATAT	AGGATTGAC	
			CGCTTAGATGAAGTACAGATTGTAACAAGT	TCGGAAGTA	
NZ_CTJD01000013.1	NZ_CTJD01000001.1	*Streptococcus agalactiae* strain GB-NI-015	TGAAAACAAGCGCAAAGCTGTCAGAAAACA	CGGAACTAA	None
			TACTTGACGAATTGAAGATGACGGAATTTA	TGGCTCCAC	
NZ_CPZS01000003.1	NZ_CPZS01000001.1	*Streptococcus agalactiae* strain IT-NI-009	TATTTGATAGCGGTAACGGGTCATATACAA	AGGCATCTA	None
			ACCTTGCTCCGATGACACCATCGCGAACCT	TGGTCTAAT	
NZ_CPVQ01000026.1	NZ_CPVQ01000002.1	*Streptococcus agalactiae* strain RBH12	AACACAGCTTCCTCGAAAGGGATATATCTA	CGGACAACT	None
NDGB01000049.1	NDGB01000023.1	*Streptococcus agalactiae* strain ST 618	ATTAAGTTGCTTAGTGCTTTCATAATCATC	TGGAATAAC	None
NDGB01000030.1			ATTAAGTTGCTTAGTGCTTTCATAATCATC	TGGAATAAC	
NZ_KQ969340.1	NZ_KQ969342.1	*Streptococcus oralis* strain DD14	TTCCATTTCTGATTTGATTCAACAGCAGCA	GGAAATCCT	None
			TACAGCGGATACAACCCCACCAATAGCCTC	AGGAATTGC	
NZ_KQ961462.1	NZ_KQ961485.1	*Streptococcus pasteurianus* strain GED7275A	TTTATTCGGCATCGGCTGGTGTTATGGACT	TGGCTGCGG	None

NZ_AWTL01000007.1	NZ_AWTL01000011.1	*Streptococcus pyogenes* GA03805	TAGAGTAAACCGAATCTTTGCCATCTCTGG	CAGTTTGAC	None
NZ_LRGN01000012.1	NZ_LRGN01000001.1	*Streptococcus pyogenes* strain SST2091-1	TAGAGTAAACCGAATCTTTGCCATCTCTGG	CAGTTTGAC	None
LRGT01000330.1	LRGT01000062.1	*Streptococcus pyogenes* strain SST2097-1	TGGTCTAACTGCGTCTGGTCTGTGAATGA	TAGGTACAA	None
NC_021838.1	NC_021838.1	*Listeria monocytogenes* strain R2-502	GGTAAAACAAGCATCGGCGAAGCAGTAACA	TGGCTTCTT	AcrIIA3 (WP_023553812.1), AcrIIA2 (WP_023553814.1), AcrIIA1 (WP_003722518.1), AcrIIA1 (WP_012581438.1)
			GGTAAAACAAGCATCGGCGAAGCAGTAACA	TGGCTACTC	
			TAGGTTTAGGGAGTAAATTAGCTCCTTTGG	CAGCTGGGT	
			TAACTTTAGATACTGCTAAAGAATTAGCAA	TGGTGCAAA	
			TTGGGCAAAATGACCGTAATAAATCCATTC	CGGTTCATC	
			TAGGTTTAGGGAGTAAATTAGCTCCTTTGG	CGGCTGGAT	

Table S11. Top Acr gene candidates within each genome ranked by AcRanker. The proteins found within the prophages of 20 *Streptococcus* genomes were ranked using AcRanker; up to the top 10 highest ranking genes are listed in ascending order. Known Acr genes and the 10 genes synthesized for biochemical testing are indicated in the rightmost column. Genomes with fewer than 10 listed have very few annotated proteins found within predicted prophages.

Organism	**Source Contig**	**Protein**	**Rank**	**Candidate # or Acr**
Streptococcus iniae strain UEL-Si1	NZ_MNAC01000021.1	WP_071127623.1	1	ML1
	NZ_MNAC01000023.1	WP_071127667.1	2	
		WP_071127683.1	3	
		WP_071127693.1	4	
	NZ_MNAC01000021.1	WP_071127625.1	5	AcrIIA3
		WP_071127624.1	6	
	NZ_MNAC01000023.1	WP_071127689.1	7	
	NZ_MNAC01000021.1	WP_071127610.1	8	
	NZ_MNAC01000023.1	WP_071127674.1	9	
	NZ_MNAC01000021.1	WP_071127619.1	10	
Streptococcus pyogenes strain SS1447	NZ_JRLL01000026.1	WP_032460883.1	1	
		WP_029713970.1	2	
		WP_003057301.1	3	
	NZ_JRLL01000072.1	WP_032461152.1	4	
		WP_076634198.1	5	
	NZ_JRLL01000026.1	WP_032460878.1	6	
	NZ_JRLL01000072.1	WP_002986828.1	7	
	NZ_JRLL01000026.1	WP_080286986.1	8	
		WP_012678849.1	9	
		WP_032460877.1	10	
Streptococcus agalactiae strain DE-NI-032	NZ_CQAV01000003.1	WP_000640620.1	1	
		WP_000164461.1	2	
		WP_025194532.1	3	
		WP_017827941.1	4	
		WP_050201842.1	5	
		WP_050305756.1	6	
		WP_001162136.1	7	

		WP_000431575.1	8	
		WP_000138374.1	9	
		WP_001872365.1	10	
Streptococcus pyogenes strain NGAS322	NZ_CP010449.1	WP_002983328.1	1	
		WP_080370149.1	2	
		WP_002983750.1	3	
		WP_002984315.1	4	
		WP_032465789.1	5	
		WP_002982773.1	6	
		WP_011054546.1	7	
		WP_010921912.1	8	
		WP_080370134.1	9	
		WP_053308468.1	10	
Streptococcus agalactiae strain CCUG 37430	NZ_ALQN01000018.1	WP_000649300.1	1	
		WP_079261174.1	2	
		WP_000660740.1	3	
		WP_000076700.1	4	
		WP_000033707.1	5	
		WP_000343312.1	6	
		WP_000130090.1	7	
		WP_000582684.1	8	
		WP_000431581.1	9	
		WP_000323860.1	10	
Streptococcus agalactiae strain DE-NI-007	NZ_CPVL01000019.1	WP_000694571.1	1	
		WP_001166092.1	2	
		WP_000359663.1	3	
		WP_000141918.1	4	
		WP_000648623.1	5	
		WP_079260963.1	6	
		WP_000205000.1	7	
		WP_000130289.1	8	
		WP_000946250.1	9	ML4
		WP_001021397.1	10	
		WP_001080841.1	12	ML5
Streptococcus agalactiae FSL S3-586	NZ_ANCM01000028.1	WP_017643458.1	1	
		WP_000134940.1	2	

	NZ_ANCM01000006.1	WP_001875290.1	3	
	NZ_ANCM01000028.1	WP_000789102.1	4	
		WP_003051787.1	5	
		WP_000032136.1	6	
		WP_000342242.1	7	
		WP_000686776.1	8	
		WP_000988928.1	9	
	NZ_ANCM01000101.1	WP_017643459.1	10	
Streptococcus agalactiae strain GB00548	NZ_ALTM01000002.1	WP_000331953.1	1	
		WP_000259017.1	2	
		WP_000793595.1	3	
		WP_079254676.1	4	
		WP_000384271.1	5	ML8
		WP_001018249.1	6	
		WP_000568029.1	7	
		WP_001097380.1	8	
		WP_001867157.1	9	
		WP_000656477.1	10	
		WP_000134666.1	12	ML9
		WP_000591144.1	29	ML7
Streptococcus agalactiae strain GB00984	NZ_ALUZ01000056.1	WP_000660738.1	1	
		WP_000164461.1	2	
		WP_017827941.1	3	
		WP_000965653.1	4	
		WP_000431574.1	5	
		WP_000138374.1	6	
		WP_000614971.1	7	
		WP_000258802.1	8	
		WP_000763911.1	9	
		WP_000118546.1	10	
Streptococcus agalactiae strain BSU188	NZ_ALRF01000068.1	WP_001042289.1	1	
		WP_000965633.1	2	ML6
		WP_025194532.1	3	
		WP_000660741.1	4	
		WP_001162136.1	5	
		WP_000274022.1	6	

		WP_000076712.1	7	
		WP_001183891.1	8	
		WP_000431576.1	9	
		WP_000763914.1	10	
Streptococcus agalactiae strain MRI Z1-022	NZ_ANEM01000074.1	WP_017648179.1	1	
		WP_079265830.1	2	
		WP_000033707.1	3	
		WP_000582684.1	4	
		WP_017648175.1	5	
		WP_017648177.1	6	
		WP_000802599.1	7	
		WP_000343901.1	8	
		WP_025195242.1	9	
		WP_000142566.1	10	
Streptococcus agalactiae strain DK-NI-014	NZ_CQEL01000002.1	WP_000421991.1	1	
		WP_000640620.1	2	
		WP_011058321.1	3	ML10
		WP_000965642.1	4	
		WP_000660741.1	5	
		WP_000906736.1	6	
		WP_001162136.1	7	
		WP_000076715.1	8	
		WP_000027835.1	9	
		WP_001872365.1	10	
Streptococcus agalactiae strain DK-PW-096	NZ_MAWX01000026.1	WP_000258802.1	1	
		WP_001229661.1	2	
		WP_001921522.1	3	
		WP_000774601.1	4	
		WP_011324937.1	5	
		WP_000218309.1	6	
		WP_079261306.1	7	
		WP_000411527.1	8	
		WP_001270064.1	9	
		WP_000659174.1	10	
Streptococcus agalactiae strain GB-NI-015	NZ_CTJD01000013.1	WP_000640620.1	1	
		WP_000660738.1	2	

		WP_000164461.1	3	
		WP_017827941.1	4	
		WP_000965655.1	5	
		WP_000431574.1	6	
		WP_000138374.1	7	
		WP_001872365.1	8	
		WP_000614971.1	9	
		WP_000258802.1	10	
Streptococcus agalactiae strain IT-NI-009	NZ_CPZS01000003.1	WP_000640620.1	1	
		WP_000164461.1	2	
		WP_079261174.1	3	
		WP_050201842.1	4	
		WP_001162136.1	5	
		WP_000431575.1	6	
		WP_000138374.1	7	
		WP_001872365.1	8	
		WP_000474006.1	9	
		WP_000258802.1	10	
Streptococcus agalactiae strain RBH12	NZ_CPVQ01000026.1	WP_000650503.1	1	
		WP_000164461.1	2	
		WP_079261174.1	3	
		WP_050198474.1	4	
		WP_001058281.1	5	
		WP_079454162.1	6	
		WP_050199334.1	7	
		WP_000612386.1	8	
		WP_000963485.1	9	
		WP_000206191.1	10	
Streptococcus agalactiae strain ST 618	NDGB01000030.1	OTG45472.1	1	
		OTG45475.1	2	
		OTG45496.1	3	
		OTG45484.1	4	
		OTG45499.1	5	
		OTG45477.1	6	
		OTG45479.1	7	
		OTG45483.1	8	

		OTG45481.1	9	
		OTG45480.1	10	
Streptococcus oralis strain DD14	NZ_KQ969340.1	WP_061420077.1	1	
		WP_061420097.1	2	
		WP_061420111.1	3	
		WP_061420115.1	4	
		WP_061420334.1	5	
		WP_061420080.1	6	
		WP_061420123.1	7	
		WP_061420133.1	8	
		WP_061420073.1	9	
		WP_061420062.1	10	
Streptococcus pasteurianus strain GED7275A	NZ_KQ961462.1	WP_061100257.1	1	
		WP_061100237.1	2	
		WP_061100224.1	3	
		WP_061100243.1	4	
		WP_061100244.1	5	
		WP_061100249.1	6	
		WP_082731474.1	7	
		WP_061100238.1	8	
		WP_061100250.1	9	
		WP_061100233.1	10	
Streptococcus pyogenes GA03805	NZ_AWTL01000007.1	WP_011528797.1	1	
		WP_011888786.1	2	
		WP_023079933.1	3	
		WP_023079900.1	4	
		WP_023079918.1	5	
		WP_002985387.1	6	
		WP_011528776.1	7	
		WP_023079897.1	8	
		WP_011017565.1	9	
		WP_023079923.1	10	
Streptococcus pyogenes strain SST2091-1	NZ_LRGN01000012.1	WP_011889039.1	1	
		WP_011285632.1	2	
		WP_010922455.1	3	
		WP_010922464.1	4	

		WP_002994106.1	5	
		WP_002994744.1	6	
		WP_063629031.1	7	
		WP_080464960.1	8	
		WP_063629030.1	9	
		WP_063629029.1	10	
Streptococcus pyogenes strain SST2097-1	LRGT01000330.1	OAC70929.1	1	
		OAC70939.1	2	
		OAC70918.1	3	
		OAC70933.1	4	
		OAC70928.1	5	
		OAC70915.1	6	
		OAC70921.1	7	
		OAC70941.1	8	
		OAC70938.1	9	
		OAC70937.1	10	
Listeria monocytogenes strain R2-502	NC_021838.1	WP_003731672.1	1	
		WP_003733710.1	2	
		WP_003731277.1	3	ML2
		WP_003731276.1	4	ML3
		WP_023553812.1	5	AcrIIA3
		WP_014601509.1	6	
		WP_003733721.1	7	
		WP_003731655.1	8	
		WP_003725074.1	9	
		WP_014601388.1	10	
		WP_023553814.1	34	AcrIIA2
		WP_003722518.1	71	AcrIIA1
		WP_012581438.1	95	AcrIIA1

Table S12. BLAST vs. AcRanker rankings for the selection candidates ML1-ML10. After selecting the 10 candidate proteins for biochemical investigation, we performed a blastp ranking to determine the ability of BLAST to predict new Acr proteins. The three validated anti-CRISPRs are indicated with tan shading and in all three cases, AcRanker gives a much higher ranking than BLAST.

Candidate	Prophage proteome size	Blastp rank (e-value)	AcRanker rank
ML1 (AcrIIA20)	56	12 (0.38)	1
ML2	190	155 (4.85)	1
ML3 (AcrIIA12)	190	132 (2.48)	2
ML4	26	4 (0.16)	9
ML5	26	16 (0.7)	12
ML6	75	37 (1.3)	2
ML7	32	29 (5.84)	29
ML8 (AcrIIA21)	32	27 (3.24)	5
ML9	11	11 (0.5)	12
ML10	74	74 (4.5)	3

References

1. Lundberg,S.M. and Lee,S.-I. (2017) A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*.
2. Shen,J., Zhang,J., Luo,X., Zhu,W., Yu,K., Chen,K., Li,Y. and Jiang,H. (2007) Predicting protein-protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. USA*, **104**, 4337–4341.
3. Pawluk,A., Bondy-Denomy,J., Cheung,V.H.W., Maxwell,K.L. and Davidson,A.R. (2014) A New Group of Phage Anti-CRISPR Genes Inhibits the Type I-E CRISPR-Cas System of Pseudomonas aeruginosa. *mBio*, **5**, e00896-14.
4. Bondy-Denomy,J., Pawluk,A., Maxwell,K.L. and Davidson,A.R. (2013) Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature*, **493**, 429–432.
5. Pawluk,A., Staals,R.H.J., Taylor,C., Watson,B.N.J., Saha,S., Fineran,P.C., Maxwell,K.L. and Davidson,A.R. (2016) Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species. *Nat. Microbiol.*, **1**, 16085.
6. Rauch,B.J., Silvis,M.R., Hultquist,J.F., Waters,C.S., McGregor,M.J., Krogan,N.J. and Bondy-Denomy,J. (2017) Inhibition of CRISPR-Cas9 with Bacteriophage Proteins. *Cell*, **168**, 150–158.
7. Hynes,A.P., Rousseau,G.M., Lemay,M.L., Horvath,P., Romero,D.A., Fremaux,C. and Moineau,S. (2017) An anti-CRISPR from a virulent streptococcal phage inhibits Streptococcus pyogenes Cas9. *Nat. Microbiol.*, **2**, 1374–1380.
8. Marino,N.D., Zhang,J.Y., Borges,A.L., Sousa,A.A., Leon,L.M., Rauch,B.J., Walton,R.T., Berry,J.D., Joung,J.K., Kleinstiver,B.P., *et al.* (2018) Discovery of widespread type I and type V CRISPR-Cas inhibitors. *Science*, **362**, 240–242.
9. Hynes,A.P., Rousseau,G.M., Agudelo,D., Goulet,A., Amigues,B., Loehr,J., Romero,D.A., Fremaux,C., Horvath,P., Doyon,Y., *et al.* (2018) Widespread anti-CRISPR proteins in virulent bacteriophages inhibit a range of Cas9 proteins. *Nat. Commun.*, **9**, 2919.
10. Uribe,R. V, van der Helm,E., Misiakou,M.A., Lee,S.W., Kol,S. and Sommer,M.O.A. (2019) Discovery and Characterization of Cas9 Inhibitors Disseminated across Seven Bacterial Phyla. *Cell Host and Microbe*, **25**, 233–241.
11. Watters,K.E., Shivram,H., Fellmann,C., Lew,R.J., McMahon,B. and Doudna,J.A. (2020) Potent CRISPR-Cas9 inhibitors from Staphylococcus genomes. *Proc. Natl. Acad. Sci. USA*, **117**, 1–9.
12. Lee,J., Mir,A., Edraki,A., Garcia,B., Amrani,N., Lou,H.E., Gainetdinov,I., Pawluk,A., Ibraheim,R., Gao,X.D., *et al.* (2018) Potent Cas9 inhibition in bacterial and human cells by AcrIIC4 and AcrIIC5 anti-CRISPR proteins. *mBio*, **9**, 1–17.
13. Bhoobalan-Chitty,Y., Johansen,T.B., Di Cianni,N. and Peng,X. (2019) Inhibition of Type III CRISPR-Cas Immunity by an Archaeal Virus-Encoded Anti-CRISPR Protein. *Cell*, **179**, 448-458.e11.
14. Kyle E. Watters, Christof Fellmann, Hua B. Bai, Shawn M. Ren and Jennifer A. Doudna (2018) Systematic discovery of natural CRISPR-Cas12a inhibitors. *Science*, **362**, 236–239.